Morphological Productivity: Corpus-Based Approaches

Marco Baroni

University of Bologna

Granada "Morphology and Corpora" Seminar

Outline

Attested and possible words

- Morphology is about defining what is a *possible word* (and explaining why it is possible)

Attested and possible words

- Morphology is about defining what is a *possible word* (and explaining why it is possible)
- *Attested* words are subset of *possible* words

Attested and possible words

- Morphology is about defining what is a *possible word* (and explaining why it is possible)
- *Attested* words are subset of *possible* words
- Need to delimit set of *possible* but *unattested* words

Word-formation and productivity

- Possible unattested words are (mostly) derived by word-formation processes (rules/schemas/. . .)

Word-formation and productivity

- Possible unattested words are (mostly) derived by word-formation processes (rules/schemas/. . .)
- However, not all wf processes are (equally) available to speakers

Word-formation and productivity

- Possible unattested words are (mostly) derived by word-formation processes (rules/schemas/. . .)
- However, not all wf processes are (equally) available to speakers
- E.g.:

Word-formation and productivity

- Possible unattested words are (mostly) derived by word-formation processes (rules/schemas/...)
- However, not all wf processes are (equally) available to speakers
- E.g.:
  - *-ness* vs. *-ity* vs. *-th*

Word-formation and productivity

- Possible unattested words are (mostly) derived by word-formation processes (rules/schemas/. . .)
- However, not all wf processes are (equally) available to speakers
- E.g.:
  - *-ness* vs. *-ity* vs. *-th*
  - NN compounding in Germanic vs. Romance languages

Word-formation and productivity

- Possible unattested words are (mostly) derived by word-formation processes (rules/schemas/...)
- However, not all wf processes are (equally) available to speakers
- E.g.:
  - *-ness* vs. *-ity* vs. *-th*
  - NN compounding in Germanic vs. Romance languages
- *-ness* and Germanic NN compounding are *productive* processes

The centrality of productivity in morphology

- Objective measure of productivity needed

The centrality of productivity in morphology

- Objective measure of productivity needed
- because any theory of morphology/word formation must:

The centrality of productivity in morphology

- Objective measure of productivity needed
- because any theory of morphology/word formation must:
  - focus on productive processes

The centrality of productivity in morphology

- Objective measure of productivity needed
- because any theory of morphology/word formation must:
  - focus on productive processes
  - explain why only certain processes are productive

The centrality of productivity in morphology

- Objective measure of productivity needed
- because any theory of morphology/word formation must:
  - focus on productive processes
  - explain why only certain processes are productive
- Vast literature on productivity (see refs.)

Productivity: the classic definition

Schultink (1961), translated by Booij

Productivity as morphological phenomenon is the possibility which language users have to form an in principle uncountable number of new words unintentionally, by means of a morphological process which is the basis of the form-meaning correspondence of some words they know.

Some issues

- "Unintentionally"

Some issues

- "Unintentionally"
- "In principle uncountable" (the *step-* problem)

Some issues

- "Unintentionally"
- "In principle uncountable" (the *step-* problem)
- How do we find out if a process can form an "in principle uncountable" number of new words?

Some issues

- "Unintentionally"
- "In principle uncountable" (the *step-* problem)
- How do we find out if a process can form an "in principle uncountable" number of new words?
- Is productivity an all-or-nothing phenomenon? Does the rate at which different productive processes grows towards uncountably many forms matter?

Some issues

- "Unintentionally"
- "In principle uncountable" (the *step*- problem)
- How do we find out if a process can form an "in principle uncountable" number of new words?
- Is productivity an all-or-nothing phenomenon? Does the rate at which different productive processes grows towards uncountably many forms matter?
- How should productivity be interpreted? Is productivity an inherent property of a process, or an epiphenomenon? (Plag 1998)

Some issues

- "Unintentionally"
- "In principle uncountable" (the *step-* problem)
- How do we find out if a process can form an "in principle uncountable" number of new words?
- Is productivity an all-or-nothing phenomenon? Does the rate at which different productive processes grows towards uncountably many forms matter?
- How should productivity be interpreted? Is productivity an inherent property of a process, or an epiphenomenon? (Plag 1998)

Productivity as all-or-nothing

- Availability (Bauer 2001): *-ness* is available, *-th* is not

Productivity as all-or-nothing

- Availability (Bauer 2001): *-ness* is available, *-th* is not
- Different from "grammatical" vs. "ungrammatical": *kingdom*, *growth* are "grammatical"

Problems with the availability approach

- Common expressions like "very productive", "marginally productive" betray shared intuition that productivity is not "black or white"

Problems with the availability approach

- Common expressions like "very productive", "marginally productive" betray shared intuition that productivity is not "black or white"
- *re-* is more productive than *de-*, but *de-* is more productive than *be-*

Problems with the availability approach

- Common expressions like "very productive", "marginally productive" betray shared intuition that productivity is not "black or white"
- *re-* is more productive than *de-*, but *de-* is more productive than *be-*
- Once available always available?

Problems with the availability approach

- Common expressions like "very productive", "marginally productive" betray shared intuition that productivity is not "black or white"
- *re-* is more productive than *de-*, but *de-* is more productive than *be-*
- Once available always available?
- Baayen (2003) finds productive uses of *-th* on the Net ("Maintainance (sic) of greenth")

Measuring (degrees of) productivity

- How do we measure how many words *could* be generated by a process, when the words that have already been generated are all we can see?

Measuring (degrees of) productivity

- How do we measure how many words *could* be generated by a process, when the words that have already been generated are all we can see?
- Early proposals (e.g., Aronoff 1976) have operationalization problems

Measuring (degrees of) productivity

- How do we measure how many words *could* be generated by a process, when the words that have already been generated are all we can see?
- Early proposals (e.g., Aronoff 1976) have operationalization problems
- Baayen and colleagues (see refs.) ground study of productivity in corpora and tradition of lexical statistics

Measuring (degrees of) productivity

- How do we measure how many words *could* be generated by a process, when the words that have already been generated are all we can see?
- Early proposals (e.g., Aronoff 1976) have operationalization problems
- Baayen and colleagues (see refs.) ground study of productivity in corpora and tradition of lexical statistics
- Corpora: you need to count something, to find out that X is more productive than Y (dictionary entries not appropriate)

Measuring (degrees of) productivity

- How do we measure how many words *could* be generated by a process, when the words that have already been generated are all we can see?
- Early proposals (e.g., Aronoff 1976) have operationalization problems
- Baayen and colleagues (see refs.) ground study of productivity in corpora and tradition of lexical statistics
- Corpora: you need to count something, to find out that X is more productive than Y (dictionary entries not appropriate)
- Lexical statistics: we must count properties of our *sample* (instances of wf process attested in the corpus) to infer properties of the *population* our sample is taken from (all possible instances of wf process)

Outline

1. Introduction
2. Quantitative productivity: Baayen's approach
3. Methodological issues in measuring quantitative productivity
4. The interpretation of (quantitative) productivity
5. Conclusion

Lexical statistics

Zipf 1949/1961, Baayen 2001, Evert 2005

- Comparison of vocabulary size and other measures of lexical richness

Lexical statistics

Zipf 1949/1961, Baayen 2001, Evert 2005

- Comparison of vocabulary size and other measures of lexical richness
- E.g., for stylometry (does Joyce use richer vocabulary than H. James?), language acquisition (how many words do 7-year old know? is the L2 learners' vocabulary significantly smaller than the one of natives?), genre/register analysis (is spoken English lexically poorer than written English)

Lexical statistics

Zipf 1949/1961, Baayen 2001, Evert 2005

- Comparison of vocabulary size and other measures of lexical richness
- E.g., for stylometry (does Joyce use richer vocabulary than H. James?), language acquisition (how many words do 7-year old know? is the L2 learners' vocabulary significantly smaller than the one of natives?), genre/register analysis (is spoken English lexically poorer than written English)
- Productivity as a nuisance: target sample (text, corpus) does not contain full vocabulary

Lexical statistics

Zipf 1949/1961, Baayen 2001, Evert 2005

- Comparison of vocabulary size and other measures of lexical richness
- E.g., for stylometry (does Joyce use richer vocabulary than H. James?), language acquisition (how many words do 7-year old know? is the L2 learners' vocabulary significantly smaller than the one of natives?), genre/register analysis (is spoken English lexically poorer than written English)
- Productivity as a nuisance: target sample (text, corpus) does not contain full vocabulary
- Development of methods to assess "growth rate" of vocabulary and estimate vocabulary size (and other measures) in whole population

Basic terminology

- N: sample/corpus size, number of *tokens* in the sample

Basic terminology

- N: sample/corpus size, number of *tokens* in the sample
- V: vocabulary size, number of distinct *types* in the sample

Basic terminology

- N: sample/corpus size, number of *tokens* in the sample
- V: vocabulary size, number of distinct *types* in the sample
- V1: *hapax legomena* count, number of word types that occur only once in the sample (for hapaxes, Count(*types*) $=$ Count(*tokens*))

Basic terminology

- N: sample/corpus size, number of *tokens* in the sample
- V: vocabulary size, number of distinct *types* in the sample
- V1: *hapax legomena* count, number of word types that occur only once in the sample (for hapaxes, Count(*types*) $=$ Count(*tokens*))
- A sample: `a b b c a a b a`

Basic terminology

- N: sample/corpus size, number of *tokens* in the sample
- V: vocabulary size, number of distinct *types* in the sample
- V1: *hapax legomena* count, number of word types that occur only once in the sample (for hapaxes, Count(*types*) = Count(*tokens*))
- A sample: `a b b c a a b a`
- N: 8; V: 3; V1: 1

Vocabulary growth curve

- The sample: `a b b c a a b a`

Vocabulary growth curve

- The sample: `a b b c a a b a`
- N: 1, V: 1, V1: 1

Vocabulary growth curve

- The sample: a b b c a a b a
- N: 1, V: 1, V1: 1
- N: 3, V: 2, V1: 1

Vocabulary growth curve

- The sample: `a b b c a a b a`
- N: 1, V: 1, V1: 1
- N: 3, V: 2, V1: 1
- N: 5, V: 3, V1: 1

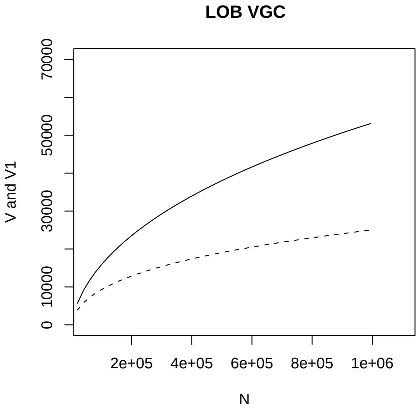
Vocabulary growth curve

- The sample: a b b c a a b a
- N: 1, V: 1, V1: 1
- N: 3, V: 2, V1: 1
- N: 5, V: 3, V1: 1
- N: 8, V: 3, V1: 1

Vocabulary growth curve

- The sample: `a b b c a a b a`
- N: 1, V: 1, V1: 1
- N: 3, V: 2, V1: 1
- N: 5, V: 3, V1: 1
- N: 8, V: 3, V1: 1
- (Most VGCs below smoothed with *binomial interpolation*)

Vocabulary growth curve of LOB corpus



Frequency spectrum

- The sample: `a b b c a a b a d`

Frequency spectrum

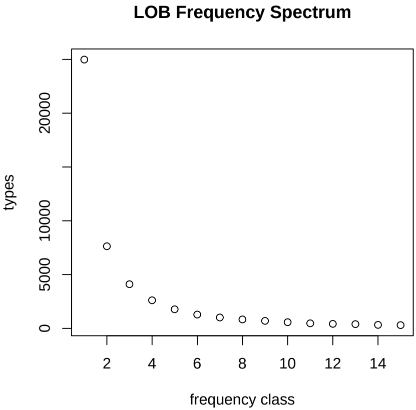
- The sample: `a b b c a a b a d`
- Frequency classes: 1 (`c, d`), 3 (`b`), 4 (`a`)

Frequency spectrum

- The sample: `a b b c a a b a d`
- Frequency classes: 1 (`c, d`), 3 (`b`), 4 (`a`)
- Frequency spectrum:

m	V(m)
1	2
3	1
4	1

Frequency spectrum of LOB corpus



Morphology, productivity and lexical statistics

- N: number of tokens characterized by target wf process in corpus

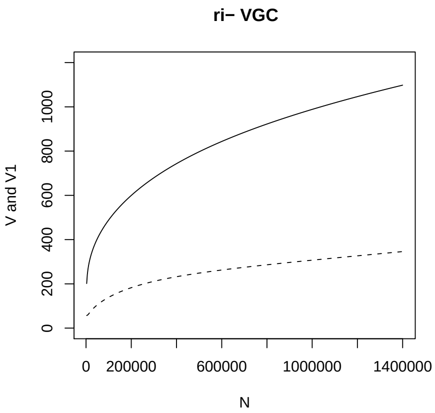
Morphology, productivity and lexical statistics

- N: number of tokens characterized by target wf process in corpus
- V: number distinct types characterized by target wf process

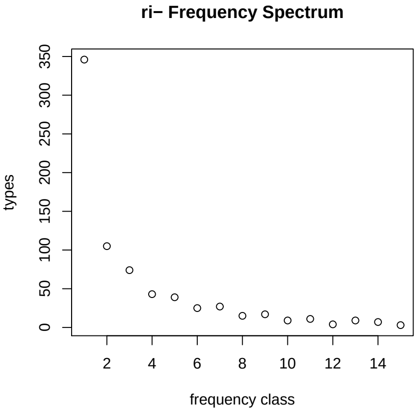
Morphology, productivity and lexical statistics

- N: number of tokens characterized by target wf process in corpus
- V: number distinct types characterized by target wf process
- V1: number of hapax legomena characterized by target wf process

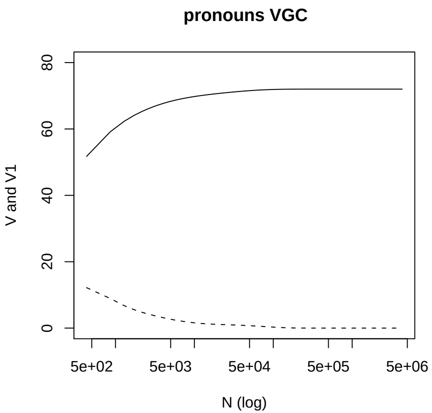
ri- in Italian *la Repubblica* corpus



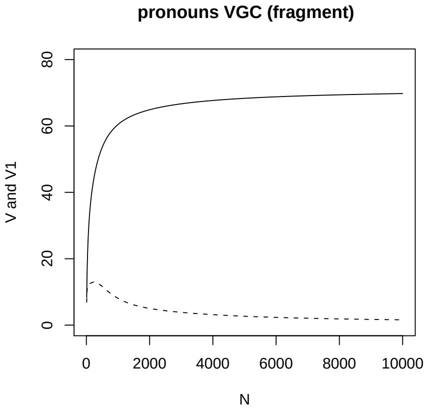
ri- in Italian *la Repubblica* corpus



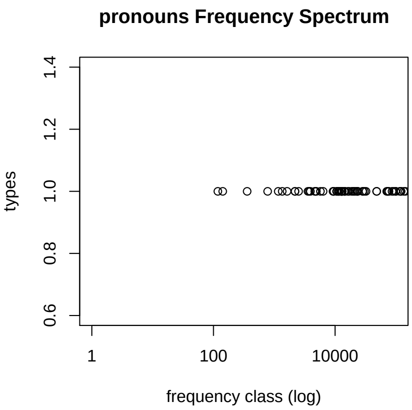
Pronouns in Italian *la Repubblica* corpus



Pronouns in Italian *la Repubblica* corpus



Pronouns in Italian *la Repubblica* corpus



V as a measure of productivity

- Valid only for corpora/samples of equal size!

V as a measure of productivity

- Valid only for corpora/samples of equal size!
- Good first approximation, but it is measuring attestedness, not potential:

V as a measure of productivity

- Valid only for corpora/samples of equal size!
- Good first approximation, but it is measuring attestedness, not potential:
  - (According to rough BNC counts) *de-* verbs have V of 141, *un-* verbs have V of 119, contra our intuition

V as a measure of productivity

- Valid only for corpora/samples of equal size!
- Good first approximation, but it is measuring attestedness, not potential:
  - (According to rough BNC counts) *de-* verbs have V of 141, *un-* verbs have V of 119, contra our intuition
  - We want productivity index of pronouns to be 0, not 72!

V as a measure of productivity

- Valid only for corpora/samples of equal size!
- Good first approximation, but it is measuring attestedness, not potential:
  - (According to rough BNC counts) *de-* verbs have V of 141, *un-* verbs have V of 119, contra our intuition
  - We want productivity index of pronouns to be 0, not 72!
- (V of whole population *could* measure potential – see below)

Hapax legomena and productivity

- Plots show relation between productivity and hapax legomena

Hapax legomena and productivity

- Plots show relation between productivity and hapax legomena
- Intuition: hapax legomena are words we did not see before in our sample, until the moment in which we sample them

Hapax legomena and productivity

- Plots show relation between productivity and hapax legomena
- Intuition: hapax legomena are words we did not see before in our sample, until the moment in which we sample them
- There is a close relation between hapax legomena and words-yet-to-be-seen

Hapax legomena and productivity

- If the word we sample after seeing N tokens is a hapax legomenon, this means that the word was not in the N tokens seen up to that point, i.e., it is a *new* word at that point

Hapax legomena and productivity

- If the word we sample after seeing N tokens is a hapax legomenon, this means that the word was not in the N tokens seen up to that point, i.e., it is a *new* word at that point
- A *productive* process is a process that is more likely than others to produce new words

Hapax legomena and productivity

- If the word we sample after seeing N tokens is a hapax legomenon, this means that the word was not in the N tokens seen up to that point, i.e., it is a *new* word at that point
- A *productive* process is a process that is more likely than others to produce new words
- Thus, the more a process is productive, the more it is likely that the next word we see that has been generated by that process is a new word

Baayen's $\mathscr{P}$

- Operationalize *productivity* of a process as probability that the next token created by the process that we sample is a new word

Baayen's $\mathscr{P}$

- Operationalize *productivity* of a process as probability that the next token created by the process that we sample is a new word
- This is same as probability that next token in sample is hapax legomenon

Baayen's $\mathscr{P}$

- Operationalize *productivity* of a process as probability that the next token created by the process that we sample is a new word
- This is same as probability that next token in sample is hapax legomenon
- Thus, we can estimate probability of sampling a new word as relative frequency of hapax legomena in our sample:
  $\mathscr{P} = \frac{V1}{N}$
  (where V1 and N are limited to words displaying the relevant process)

Baayen's $\mathscr{P}$

$$\mathscr{P} = \frac{V1}{N}$$

- Probability to sample token representing type we will never encounter again (token labeled "hapax") at first stage of sampling (when we are at the beginning of N-token-sample) is given by the proportion of hapaxes in the whole N-token-sample divided by the total number of tokens in the sample

Baayen's $\mathscr{P}$

$$\mathscr{P} = \frac{V1}{N}$$

- Probability to sample token representing type we will never encounter again (token labeled "hapax") at first stage of sampling (when we are at the beginning of N-token-sample) is given by the proportion of hapaxes in the whole N-token-sample divided by the total number of tokens in the sample
- Thus, this must also be probability that *last* token sampled represents new type

Baayen's $\mathscr{P}$

$$\mathscr{P} = \frac{V1}{N}$$

- Probability to sample token representing type we will never encounter again (token labeled "hapax") at first stage of sampling (when we are at the beginning of N-token-sample) is given by the proportion of hapaxes in the whole N-token-sample divided by the total number of tokens in the sample
- Thus, this must also be probability that *last* token sampled represents new type
- $\mathscr{P}$ as productivity measure matches intuition that productivity should measure *potential* of process to generate new forms

$\mathscr{P}$ as vocabulary growth rate

- $\mathscr{P}$ measures the potentiality of growth of V in a very literal way, i.e., it is the growth rate of V, the rate at which vocabulary size increases

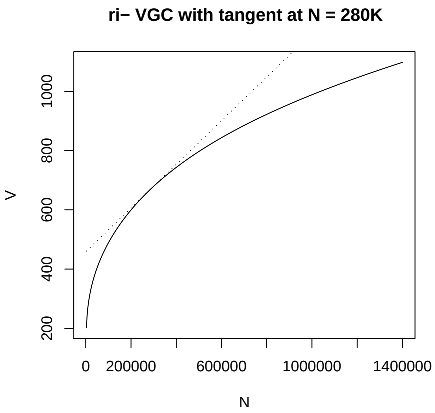
$\mathscr{P}$ as vocabulary growth rate

- $\mathscr{P}$ measures the potentiality of growth of V in a very literal way, i.e., it is the growth rate of V, the rate at which vocabulary size increases
- $\mathscr{P}$ is (approximation to) the *derivative* of V at N, i.e., the slope of the tangent to the vocabulary growth curve at N (Baayen 2001, pp. 49-50)

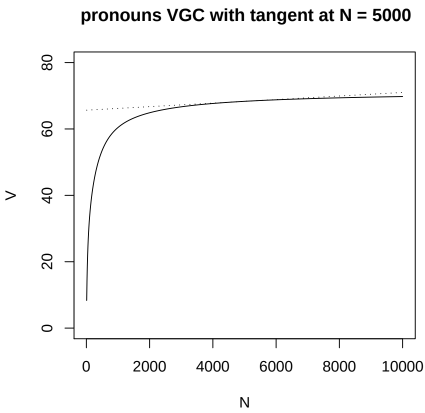
$\mathscr{P}$ as vocabulary growth rate

- $\mathscr{P}$ measures the potentiality of growth of V in a very literal way, i.e., it is the growth rate of V, the rate at which vocabulary size increases
- $\mathscr{P}$ is (approximation to) the *derivative* of V at N, i.e., the slope of the tangent to the vocabulary growth curve at N (Baayen 2001, pp. 49-50)
- Again, "rate of growth" of vocabulary generated by wf process seems good match for intuition about productivity of wf process

ri- in Italian *la Repubblica* corpus



Pronouns in Italian *la Repubblica* corpus



Baayen's $\mathscr{P}$ and intuition

class	V	V1	N	$\mathscr{P}$
it. ri-	1098	346	1,399,898	0.00025
it. pronouns	72	0	4,313,123	0
en. un-	119	25	7,618	.00328
en. de-	141	16	86,130	.000185

Applications of $\mathscr{P}$

- Extensive tradition of corpus-based analyses of derivational morphology based on $\mathscr{P}$ (and V), by Baayen and colleagues (esp. English and Dutch), but not only (e.g., Lüdeling and Evert on German morphology, Gaeta and Ricca on Italian morphology)

Applications of $\mathscr{P}$

- Extensive tradition of corpus-based analyses of derivational morphology based on $\mathscr{P}$ (and V), by Baayen and colleagues (esp. English and Dutch), but not only (e.g., Lüdeling and Evert on German morphology, Gaeta and Ricca on Italian morphology)
- $\mathscr{P}$ used as an exploratory tool

Baayen & Lieber 1991 on English derivation

- Large scale study of English derivation based on $\mathscr{P}$ and V with statistics extracted from CELEX database ($=$ 18M tokens version of COBUILD)

Baayen & Lieber 1991 on English derivation

- Large scale study of English derivation based on $\mathscr{P}$ and V with statistics extracted from CELEX database ($=$ 18M tokens version of COBUILD)
- A few results:

Baayen & Lieber 1991 on English derivation

- Large scale study of English derivation based on $\mathscr{P}$ and V with statistics extracted from CELEX database (= 18M tokens version of COBUILD)
- A few results:
  - *-ness* > *-ity*; *-ish* > *-ous*; *un-* > *in-*; *-ation* > *-al*, *-ment*

Baayen & Lieber 1991 on English derivation

- Large scale study of English derivation based on $\mathscr{P}$ and V with statistics extracted from CELEX database ($=$ 18M tokens version of COBUILD)
- A few results:
  - *-ness* $>$ *-ity*; *-ish* $>$ *-ous*; *un-* $>$ *in-*; *-ation* $>$ *-al*, *-ment*
  - …but affixes such as *-ity*, *-ous* and *in-* have $\mathscr{P} > 0$ (and there are category-of-the-base effects)

Baayen & Lieber 1991 on English derivation

- Large scale study of English derivation based on $\mathscr{P}$ and V with statistics extracted from CELEX database ($=$ 18M tokens version of COBUILD)
- A few results:
  - *-ness* $>$ *-ity*; *-ish* $>$ *-ous*; *un-* $>$ *in-*; *-ation* $>$ *-al*, *-ment*
  - ...but affixes such as *-ity*, *-ous* and *in-* have $\mathscr{P} > 0$ (and there are category-of-the-base effects)
  - dual nature of *re-*: few high frequency forms make it look unproductive (*remove, recite, recall...*) (but see below on *re-*, $\mathscr{P}$ and sample size)

Plag, Dalton-Puffer & Baayen (1999)

Productivity across speech and writing

- $\mathscr{P}$ and V in written, demographic spoken and context-governed spoken sections of BNC (keep affix constant, change corpus)

Plag, Dalton-Puffer & Baayen (1999)

Productivity across speech and writing

- $\mathscr{P}$ and V in written, demographic spoken and context-governed spoken sections of BNC (keep affix constant, change corpus)
- Strong effect of "register", with productivity higher in written than spoken, and context-governed spoken higher than demographic spoken (productivity as a dimension of register variation)

Plag, Dalton-Puffer & Baayen (1999)

Productivity across speech and writing

- $\mathscr{P}$ and V in written, demographic spoken and context-governed spoken sections of BNC (keep affix constant, change corpus)
- Strong effect of "register", with productivity higher in written than spoken, and context-governed spoken higher than demographic spoken (productivity as a dimension of register variation)
- Different productivity of different affixes in different registers: *-like* is "written-only", *-ness* is strongly "written", productivity reversal of *-ize* and *-ish* in written vs. demographic

Plag, Dalton-Puffer & Baayen (1999)

Productivity across speech and writing

- $\mathscr{P}$ and V in written, demographic spoken and context-governed spoken sections of BNC (keep affix constant, change corpus)
- Strong effect of "register", with productivity higher in written than spoken, and context-governed spoken higher than demographic spoken (productivity as a dimension of register variation)
- Different productivity of different affixes in different registers: *-like* is "written-only", *-ness* is strongly "written", productivity reversal of *-ize* and *-ish* in written vs. demographic
- Productivity is affected by register, it cannot be explained in purely structural terms

Other examples

- Gaeta and Ricca (2003): on the extremely high productivity of the Italian "neo-"prefixes *mega-, iper-, super-, maxi-, ultra-*

Other examples

- Gaeta and Ricca (2003): on the extremely high productivity of the Italian “neo-”prefixes *mega-, iper-, super-, maxi-, ultra-*
- Lüdeling and Evert (2005): medical and non-medical *-itis* in XXth century German (with focus on methodological aspects)

Outline

1. Introduction
2. Quantitative productivity: Baayen's approach
3. Methodological issues in measuring quantitative productivity
4. The interpretation of (quantitative) productivity
5. Conclusion

Methodological issues

- Pre-processing/preparing the data

Methodological issues

- Pre-processing/preparing the data
- Effect of N on $\mathscr{P}$

Pre-processing

- IT IS IMPORTANT!!!

Pre-processing

- IT IS IMPORTANT!!!
- Baayen, strangely, does not seem to worry

Pre-processing

- IT IS IMPORTANT!!!
- Baayen, strangely, does not seem to worry
- At least two aspects:

Pre-processing

- IT IS IMPORTANT!!!
- Baayen, strangely, does not seem to worry
- At least two aspects:
  - Problems of (automated) data-cleaning/complex word identification (Evert and Lüdeling 2001)

Pre-processing

- IT IS IMPORTANT!!!
- Baayen, strangely, does not seem to worry
- At least two aspects:
  - Problems of (automated) data-cleaning/complex word identification (Evert and Lüdeling 2001)
  - Theoretical issues (delimitation and identification of application of a wf process) (Gaeta and Ricca 2003, to appear)

Automated data-cleaning/complex word identification

- Often necessary (13,850 types begin with *re-* in BNC, 103,941 types begin with *ri-* in itWaC)

Automated data-cleaning/complex word identification

- Often necessary (13,850 types begin with *re-* in BNC, 103,941 types begin with *ri-* in itWaC)
- We can rely on:

Automated data-cleaning/complex word identification

- Often necessary (13,850 types begin with *re-* in BNC, 103,941 types begin with *ri-* in itWaC)
- We can rely on:
  - POS tagging

Automated data-cleaning/complex word identification

- Often necessary (13,850 types begin with *re-* in BNC, 103,941 types begin with *ri-* in itWaC)
- We can rely on:
  - POS tagging
  - Lemmatization

Automated data-cleaning/complex word identification

- Often necessary (13,850 types begin with *re-* in BNC, 103,941 types begin with *ri-* in itWaC)
- We can rely on:
  - POS tagging
  - Lemmatization
  - Pattern matching heuristics (e.g., candidate prefixed form must be analyzable as *PRE+VERB*, with VERB independently attested in corpus)

The problem with low frequency words

- Correct analysis of low frequency words is fundamental to measure productivity

The problem with low frequency words

- Correct analysis of low frequency words is fundamental to measure productivity
- However, automated tools will tend to have lowest performance on low frequency forms:

The problem with low frequency words

- Correct analysis of low frequency words is fundamental to measure productivity
- However, automated tools will tend to have lowest performance on low frequency forms:
    - Statistical tools will suffer from lack of relevant training data

The problem with low frequency words

- Correct analysis of low frequency words is fundamental to measure productivity
- However, automated tools will tend to have lowest performance on low frequency forms:
  - Statistical tools will suffer from lack of relevant training data
  - Manually-crafted tools will probably lack the relevant resources

The problem with low frequency words

- Correct analysis of low frequency words is fundamental to measure productivity
- However, automated tools will tend to have lowest performance on low frequency forms:
  - Statistical tools will suffer from lack of relevant training data
  - Manually-crafted tools will probably lack the relevant resources
- Problems in both directions (under- and overestimation of hapax counts)

The problem with low frequency words

- Correct analysis of low frequency words is fundamental to measure productivity
- However, automated tools will tend to have lowest performance on low frequency forms:
  - Statistical tools will suffer from lack of relevant training data
  - Manually-crafted tools will probably lack the relevant resources
- Problems in both directions (under- and overestimation of hapax counts)
- Part of the more general "95% performance" problem

Underestimation of hapaxes

- The Italian TreeTagger lemmatizer is lexicon-based; out-of-lexicon words (e.g., productively formed words containing a prefix) are lemmatized as UNKNOWN

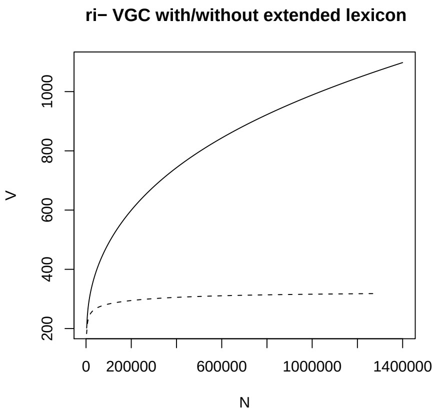
Underestimation of hapaxes

- The Italian TreeTagger lemmatizer is lexicon-based; out-of-lexicon words (e.g., productively formed words containing a prefix) are lemmatized as UNKNOWN
- No prefixed word with dash (*ri-cadere*) is in lexicon

Underestimation of hapaxes

- The Italian TreeTagger lemmatizer is lexicon-based; out-of-lexicon words (e.g., productively formed words containing a prefix) are lemmatized as UNKNOWN
- No prefixed word with dash (*ri-cadere*) is in lexicon
- Writers are more likely to use dash to mark transparent morphological structure

Productivity of *ri-* with and without an extended lexicon



Overestimation of hapaxes

- "Noise" generates hapax legomena

Overestimation of hapaxes

- "Noise" generates hapax legomena
- The Italian TreeTagger seems to think that dashed expressions containing pronoun-like strings are pronouns

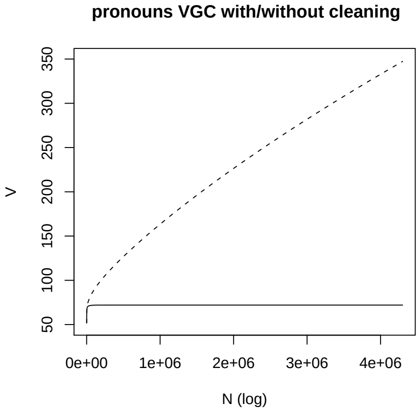
Overestimation of hapaxes

- "Noise" generates hapax legomena
- The Italian TreeTagger seems to think that dashed expressions containing pronoun-like strings are pronouns
- Dashed strings can be anything, including full sentences

Overestimation of hapaxes

- "Noise" generates hapax legomena
- The Italian TreeTagger seems to think that dashed expressions containing pronoun-like strings are pronouns
- Dashed strings can be anything, including full sentences
- This creates a lot of pseudo-pronoun hapaxes: *tu-tu, parapaponzi-ponzi-pò, altri-da-lui-simili-a-lui*

Productivity of the pronoun class before and after cleaning



$\mathscr{P}$ (and V) with/without correct post-processing

- With:

class	V	V1	N	$\mathscr{P}$
ri-	1098	346	1,399,898	0.00025
pronouns	72	0	4,313,123	0

- Without:

class	V	V1	N	$\mathscr{P}$
ri-	318	8	1,268,244	0.000006
pronouns	348	206	4,314,381	0.000048

Linguistic analysis issues

- Which of the following forms should be counted as prefixed forms with *re-*? *redo, remove, retain, remake* (as a noun), *resyllabification*

Linguistic analysis issues

- Which of the following forms should be counted as prefixed forms with *re-*? *redo, remove, retain, remake* (as a noun), *resyllabification*
- If we remove *remove* because it is lexicalized, aren't we boosting up the productivity of *re-* in a circular way?

Linguistic analysis issues

- Which of the following forms should be counted as prefixed forms with *re-*? *redo, remove, retain, remake* (as a noun), *resyllabification*
- If we remove *remove* because it is lexicalized, aren't we boosting up the productivity of *re-* in a circular way?
- Are there two *in-* prefixes? *inanimate* vs. *inchoative*

Linguistic analysis issues

- Which of the following forms should be counted as prefixed forms with *re-*? *redo, remove, retain, remake* (as a noun), *resyllabification*
- If we remove *remove* because it is lexicalized, aren't we boosting up the productivity of *re-* in a circular way?
- Are there two *in-* prefixes? *inanimate* vs. *inchoative*
- How about *re-*? *re-conquer the city* vs. *re-play the song*

Linguistic analysis issues

- Which of the following forms should be counted as prefixed forms with *re-*? *redo, remove, retain, remake* (as a noun), *resyllabification*
- If we remove *remove* because it is lexicalized, aren't we boosting up the productivity of *re-* in a circular way?
- Are there two *in-* prefixes? *inanimate* vs. *inchoative*
- How about *re-*? *re-conquer the city* vs. *re-play the song*
- Is *deXize* a different wf process from *de-*?

THESE ARE SERIOUS PROBLEMS!

- Given the current state of NLP tools (especially for languages other than English)

THESE ARE SERIOUS PROBLEMS!

- Given the current state of NLP tools (especially for languages other than English)
- and the typical resources of morphologists

THESE ARE SERIOUS PROBLEMS!

- Given the current state of NLP tools (especially for languages other than English)
- and the typical resources of morphologists
- large-scale, methodologically sound quantitative productivity studies are unfeasible

$\mathscr{P}$ and sample size

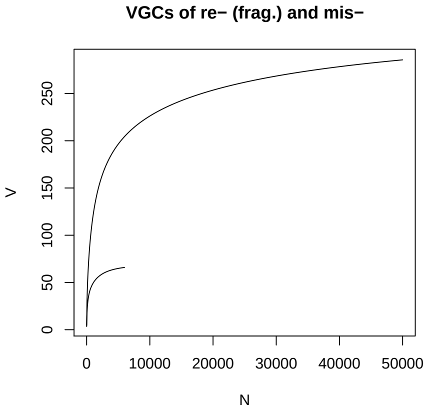
- It should be obvious that as N increases, V also increases (for at-least-mildly-productive processes)

$\mathscr{P}$ and sample size

- It should be obvious that as N increases, V also increases (for at-least-mildly-productive processes)
- Thus, V cannot be compared at different Ns

V and N

English *re-* and *mis-*



$\mathscr{P}$ and sample size

- It should be obvious that as N increases, V also increases (for at-least-mildly-productive processes)
- Thus, V cannot be compared at different Ns

$\mathscr{P}$ and sample size

- It should be obvious that as N increases, V also increases (for at-least-mildly-productive processes)
- Thus, V cannot be compared at different Ns
- However, the growth rate is also systematically decreasing as N becomes larger

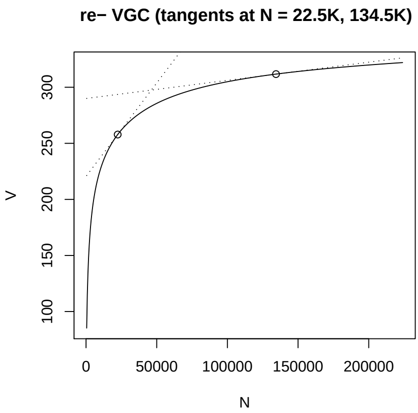
$\mathscr{P}$ and sample size

- It should be obvious that as N increases, V also increases (for at-least-mildly-productive processes)
- Thus, V cannot be compared at different Ns
- However, the growth rate is also systematically decreasing as N becomes larger
- At the beginning, any word will be a hapax legomenon; as sample increases, hapaxes will be increasingly lower proportion of sample

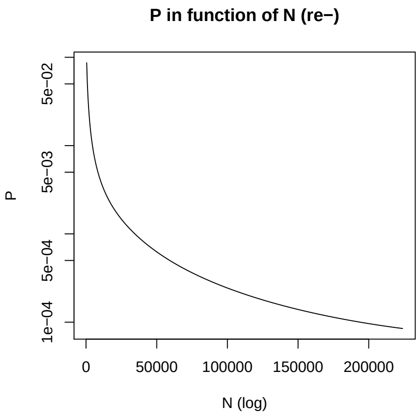
$\mathscr{P}$ and sample size

- It should be obvious that as N increases, V also increases (for at-least-mildly-productive processes)
- Thus, V cannot be compared at different Ns
- However, the growth rate is also systematically decreasing as N becomes larger
- At the beginning, any word will be a hapax legomenon; as sample increases, hapaxes will be increasingly lower proportion of sample
- A specific instance of the more general problem of "variable constants" (Tweedie and Baayen 1998) in lexical statistics (cf. type/token ratio)

Growth rate of *re-* at different sample sizes



$\mathscr{P}$ as a function of N (*re-*)



The solution: control N

Gaeta and Ricca (to appear)

- Always compute $\mathscr{P}$ at comparable N

The solution: control N

Gaeta and Ricca (to appear)

- Always compute $\mathscr{P}$ at comparable N
- Given two wf processes with sample sizes N_a and N_b, with $N_a > N_b$, measure $\mathscr{P}$ at N_b for both processes

The solution: control N

Gaeta and Ricca (to appear)

- Always compute $\mathscr{P}$ at comparable N
- Given two wf processes with sample sizes N_a and N_b, with $N_a > N_b$, measure $\mathscr{P}$ at N_b for both processes
- Denominator of $\mathscr{P} = \frac{V1}{N}$ is fixed, so this amounts to comparing V1, the number of hapax legomena

The solution: control N

Gaeta and Ricca (to appear)

- Always compute $\mathscr{P}$ at comparable N
- Given two wf processes with sample sizes N_a and N_b, with $N_a > N_b$, measure $\mathscr{P}$ at N_b for both processes
- Denominator of $\mathscr{P} = \frac{V1}{N}$ is fixed, so this amounts to comparing V1, the number of hapax legomena
- If more than 2 processes are compared, do comparison pairwise and use transitivity, to minimize data loss

The solution: control N

Gaeta and Ricca (to appear)

- Always compute $\mathscr{P}$ at comparable N
- Given two wf processes with sample sizes N_a and N_b, with $N_a > N_b$, measure $\mathscr{P}$ at N_b for both processes
- Denominator of $\mathscr{P} = \frac{V1}{N}$ is fixed, so this amounts to comparing V1, the number of hapax legomena
- If more than 2 processes are compared, do comparison pairwise and use transitivity, to minimize data loss
- I.e., if 3 processes have sample sizes $N_a > N_b > N_c$, compare processes *a* and *b* at N_b, *b* and *c* at N_c and infer productivity ranking of *a* and *c* on the basis of their relationship to *b*

Controlling N: $\mathscr{P}$

class	$N = 6097$	$N = 35107$	$N = 223970$
re-	0.007	0.00098	0.000085
en-	0.00075	0.00014	NA
mis-	0.00082	NA	NA

Controlling N: V1 (interpolated values!)

class	$N = 6097$	$N = 35107$	$N = 223970$
re-	43.7	34.4	19
en-	4.5	5	NA
mis-	5	NA	NA

Problems

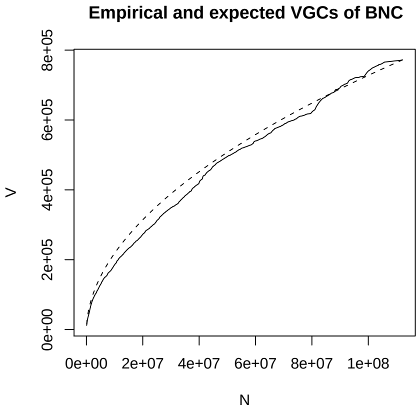
- We are throwing away data

Problems

- We are throwing away data
- Clumpiness and other non-randomness effects

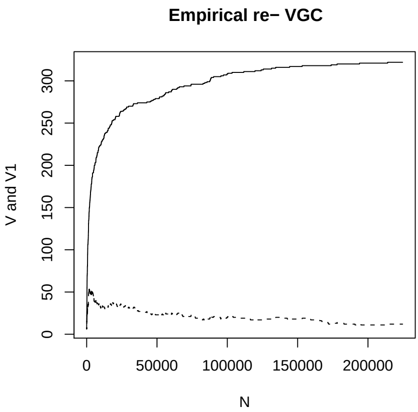
Non-randomness

Empirical and interpolated VGCs of BNC



Non-randomness

The “real” *re-* VGC



Non-randomness

- At least two issues:

Non-randomness

- At least two issues:
  - Clumpiness (well above one third of the non hapaxes in the "too much" data-set of Lüdeling/Baroni/Evert occur more than once in the same document)

Non-randomness

- At least two issues:
  - Clumpiness (well above one third of the non hapaxes in the "too much" data-set of Lüdeling/Baroni/Evert occur more than once in the same document)
  - Effects of specialized language, genre, register (in our version of BNC, spoken texts are almost entirely at end of corpus)

Non-randomness

- At least two issues:
  - Clumpiness (well above one third of the non hapaxes in the "too much" data-set of Lüdeling/Baroni/Evert occur more than once in the same document)
  - Effects of specialized language, genre, register (in our version of BNC, spoken texts are almost entirely at end of corpus)
- For less frequent process, we take sample from whole corpus, whereas for more frequent process we take sample from first N_{sub} tokens, probably resulting in more clumpiness and less variety of genre and topics

The importance of randomization

- Randomize the order of words in corpus

The importance of randomization

- Randomize the order of words in corpus
- This can be done efficiently and soundly by using *binomial interpolation* (Baayen 2001, ch. 2)

The importance of randomization

- Randomize the order of words in corpus
- This can be done efficiently and soundly by using *binomial interpolation* (Baayen 2001, ch. 2)
- Binomial interpolation produces *expected values* of V and V1 for arbitrary sample sizes ($< N$) that can be thought of as the average of an infinite number of randomizations

The importance of randomization

- Randomize the order of words in corpus
- This can be done efficiently and soundly by using *binomial interpolation* (Baayen 2001, ch. 2)
- Binomial interpolation produces *expected values* of V and V1 for arbitrary sample sizes ($< N$) that can be thought of as the average of an infinite number of randomizations
- Most plots shown on these slides are based on binomial interpolation

The importance of randomization

- Counting number of documents in which a word occurs, rather than overall occurrences, might be a cure for clumpiness (but increases data-sparseness problems, and complicates the assumptions about sampling)

The importance of randomization

- Counting number of documents in which a word occurs, rather than overall occurrences, might be a cure for clumpiness (but increases data-sparseness problems, and complicates the assumptions about sampling)
- However, non-randomized VGC plot provides very valuable information, and should always be included in quantitative productivity studies

Non-randomness: a bigger problem

- The whole corpus is probably a non-random sample of the "population" we are interested in (e.g., the population of words illustrating word formation with *re-*, or the population of words known by an English speaker)

Non-randomness: a bigger problem

- The whole corpus is probably a non-random sample of the "population" we are interested in (e.g., the population of words illustrating word formation with *re-*, or the population of words known by an English speaker)
- Unfortunately, we cannot take a randomized sub-sample from the whole population like we can do when taking a sub-sample from the whole corpus (that's what a corpus is supposed to be in the first instance!)

Non-randomness and parametric (lexico-)statistical models

- Non-randomness problems are seriously affecting the quality of parametric statistical models (Evert and Baroni to appear, and ongoing work)

Non-randomness and parametric (lexico-)statistical models

- Non-randomness problems are seriously affecting the quality of parametric statistical models (Evert and Baroni to appear, and ongoing work)
- This is a pity, since these models would allow us to *extrapolate* V and V1 to arbitrary values (including V and V1 in the whole population), and to test significance of differences

Non-randomness and parametric (lexico-)statistical models

- Non-randomness problems are seriously affecting the quality of parametric statistical models (Evert and Baroni to appear, and ongoing work)
- This is a pity, since these models would allow us to *extrapolate* V and V1 to arbitrary values (including V and V1 in the whole population), and to test significance of differences
- (Please stay tuned)

Outline

Productivity: cause or effect?

- Is productivity (as measured by $\mathscr{P}$ and related measures) a cause or an effect (an epiphenomenon)?

Productivity: cause or effect?

- Is productivity (as measured by $\mathscr{P}$ and related measures) a cause or an effect (an epiphenomenon)?
- Does $\mathscr{P}$ correspond to an "activation level" in the mind of the speaker, or should it be explained by other factors?

Productivity: cause or effect?

- Is productivity (as measured by $\mathscr{P}$ and related measures) a cause or an effect (an epiphenomenon)?
- Does $\mathscr{P}$ correspond to an "activation level" in the mind of the speaker, or should it be explained by other factors?
- If so, what kinds of factors?

Productivity: cause or effect?

- Is productivity (as measured by $\mathscr{P}$ and related measures) a cause or an effect (an epiphenomenon)?
- Does $\mathscr{P}$ correspond to an "activation level" in the mind of the speaker, or should it be explained by other factors?
- If so, what kinds of factors?
- General agreement (often implicit) is that $\mathscr{P}$ is an epiphenomenon

Productivity: cause or effect?

- Is productivity (as measured by $\mathscr{P}$ and related measures) a cause or an effect (an epiphenomenon)?
- Does $\mathscr{P}$ correspond to an "activation level" in the mind of the speaker, or should it be explained by other factors?
- If so, what kinds of factors?
- General agreement (often implicit) is that $\mathscr{P}$ is an epiphenomenon
- See, e.g., Plag (1999), who uses quantitative productivity as exploratory tool, and looks for qualitative structural explanations (phonological, semantic, morphosyntactic) of different degree of productivity of similar affixes

Corpus-based "explanations" of $\mathscr{P}$

- Hay and Baayen (2002, 2004; see also Baayen, 2003): parsing and productivity

Corpus-based “explanations” of $\mathscr{P}$

- Hay and Baayen (2002, 2004; see also Baayen, 2003): parsing and productivity
- My ongoing work on productivity and semantic transparency (Baroni and Vegnaduzzo 2003, Baroni 2005)

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:
  - E.g., *illegible* is more frequent than *legible*, and it behaves as morphologically simple

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:
  - E.g., *illegible* is more frequent than *legible*, and it behaves as morphologically simple
  - *illiberal* is less frequent that *liberal*, and it behaves as morphologically complex

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:
  - E.g., *illegible* is more frequent than *legible*, and it behaves as morphologically simple
  - *illiberal* is less frequent that *liberal*, and it behaves as morphologically complex
- Parsing explanation in a *dual route* race model – when analyzing a derived word

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:
    - E.g., *illegible* is more frequent than *legible*, and it behaves as morphologically simple
    - *illiberal* is less frequent that *liberal*, and it behaves as morphologically complex
- Parsing explanation in a *dual route* race model – when analyzing a derived word
    - If base is more frequent than derived form, it is retrieved faster, and base+affix analysis wins

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:
  - E.g., *illegible* is more frequent than *legible*, and it behaves as morphologically simple
  - *illiberal* is less frequent that *liberal*, and it behaves as morphologically complex
- Parsing explanation in a *dual route* race model – when analyzing a derived word
  - If base is more frequent than derived form, it is retrieved faster, and base+affix analysis wins
  - If derived form is more frequent, it is retrieved as a whole before base+affix analysis is accessed

Relative frequency and parsing

- One important result of Hay (2003): in a number of tasks, affixed forms behave as morphologically complex iff their base is more frequent than the affixed form itself:
  - E.g., *illegible* is more frequent than *legible*, and it behaves as morphologically simple
  - *illiberal* is less frequent that *liberal*, and it behaves as morphologically complex
- Parsing explanation in a *dual route* race model – when analyzing a derived word
  - If base is more frequent than derived form, it is retrieved faster, and base+affix analysis wins
  - If derived form is more frequent, it is retrieved as a whole before base+affix analysis is accessed
- The higher the base-to-derived-form relative frequency is, the more likely it is that a word is treated as complex

Parsing and productivity

- Hay and Baayen's hypothesis:

Parsing and productivity

- Hay and Baayen's hypothesis:
  - Affixes that appear in many words that are parsed as complex in language perception (i.e., appear in many derived words that have lower frequency than their bases) will be more "active" in lexicon

Parsing and productivity

- Hay and Baayen's hypothesis:
  - Affixes that appear in many words that are parsed as complex in language perception (i.e., appear in many derived words that have lower frequency than their bases) will be more "active" in lexicon
  - I.e., they will be more available for word formation, i.e., more productive

Parsing and productivity

- Hay and Baayen's hypothesis:
  - Affixes that appear in many words that are parsed as complex in language perception (i.e., appear in many derived words that have lower frequency than their bases) will be more "active" in lexicon
  - I.e., they will be more available for word formation, i.e., more productive
  - Predicts correlation between $\mathscr{P}$ and relative frequency

Parsing and productivity

- Hay and Baayen's hypothesis:
  - Affixes that appear in many words that are parsed as complex in language perception (i.e., appear in many derived words that have lower frequency than their bases) will be more "active" in lexicon
  - I.e., they will be more available for word formation, i.e., more productive
  - Predicts correlation between $\mathscr{P}$ and relative frequency
- Hay and Baayen (2002) report high correlation between $\mathscr{P}$ and relative frequency for 80 English derivation affixes

Problems

- Could the correlation be due to the fact that both measures heavily rely on the number of low(est) frequency forms that contain a certain affix?

Problems

- Could the correlation be due to the fact that both measures heavily rely on the number of low(est) frequency forms that contain a certain affix?
- More importantly, both $\mathscr{P}$ and relative frequency seem to be useful *indices* of parsability/productivity, effects, not causes!

Problems

- Could the correlation be due to the fact that both measures heavily rely on the number of low(est) frequency forms that contain a certain affix?
- More importantly, both $\mathscr{P}$ and relative frequency seem to be useful *indices* of parsability/productivity, effects, not causes!
- If productivity is caused by low relative frequency of bases, what causes this low relative frequency? (Or vice versa?)

Problems

- Could the correlation be due to the fact that both measures heavily rely on the number of low(est) frequency forms that contain a certain affix?
- More importantly, both $\mathscr{P}$ and relative frequency seem to be useful *indices* of parsability/productivity, effects, not causes!
- If productivity is caused by low relative frequency of bases, what causes this low relative frequency? (Or vice versa?)
- Nature of variables as epiphenomenal indices seems to be recognized by Hay and Baayen (2004), which analyze a constellation of densely inter-correlated measures related to parsability and productivity

Productivity and semantics

- Observation: we have much clearer intuitions about what productive affixes mean than about what unproductive affixes mean

Productivity and semantics

- Observation: we have much clearer intuitions about what productive affixes mean than about what unproductive affixes mean
- Cf. *redo* vs. *enlarge*

Productivity and semantics

- Observation: we have much clearer intuitions about what productive affixes mean than about what unproductive affixes mean
- Cf. *redo* vs. *enlarge*
- Hypothesis: an affix is productive as long as it has well-defined meaning in the language (it is easier to acquire the relevant semantic generalization, and thus to use it to form new words)

Productivity and semantics

- Observation: we have much clearer intuitions about what productive affixes mean than about what unproductive affixes mean
- Cf. *redo* vs. *enlarge*
- Hypothesis: an affix is productive as long as it has well-defined meaning in the language (it is easier to acquire the relevant semantic generalization, and thus to use it to form new words)
- Prediction: semantic transparency of forms containing an affix will be correlated with productivity of affix

Productivity and semantics

- Observation: we have much clearer intuitions about what productive affixes mean than about what unproductive affixes mean
- Cf. *redo* vs. *enlarge*
- Hypothesis: an affix is productive as long as it has well-defined meaning in the language (it is easier to acquire the relevant semantic generalization, and thus to use it to form new words)
- Prediction: semantic transparency of forms containing an affix will be correlated with productivity of affix
- Here, direction of causation should be clear

Measuring semantic transparency

- By hand, it is hard. . .

Measuring semantic transparency

- By hand, it is hard...
- but computational linguists have developed methods to measure semantic similarity among words (e.g., Manning and Schütze 1999)

Measuring semantic transparency

- By hand, it is hard...
- but computational linguists have developed methods to measure semantic similarity among words (e.g., Manning and Schütze 1999)
- Degree of semantic transparency of complex form is degree of semantic similarity between complex form and its base

Contextual approach to meaning

- Cruse 1986 (p. 1):

Contextual approach to meaning

- Cruse 1986 (p. 1):

> *[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts.*

Contextual approach to meaning

- Cruse 1986 (p. 1):

  > *[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts.*

- Two knowledge poor operationalizations:

Contextual approach to meaning

- Cruse 1986 (p. 1):

  > *[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts.*

- Two knowledge poor operationalizations:
  - Semantically similar words occur in similar contexts

Contextual approach to meaning

- Cruse 1986 (p. 1):

  > *[T]he semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts [...] [T]here are are good reasons for a principled limitation to linguistic contexts.*

- Two knowledge poor operationalizations:
  - Semantically similar words occur in similar contexts
  - Semantically similar words occur near each other

Contextual similarity

- Cosine (correlation) of normalized vectors representing co-occurrence frequency with all words within a certain window:

$$\cos(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$$

Contextual similarity

- Cosine (correlation) of normalized vectors representing co-occurrence frequency with all words within a certain window:

$$\cos(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i$$

- My parameters:
  - Targets: affixed form/base pairs
  - Contexts: all content words
  - Window: 1 sentence

Co-occurrence

- Measured by *Mutual Information* (MI):

$$MI(w_1, w_2) = \log \frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)}$$

Co-occurrence

- Measured by *Mutual Information* (MI):

$$MI(w_1, w_2) = \log \frac{Pr(w_1, w_2)}{Pr(w_1)Pr(w_2)}$$

- My parameters:
  - Targets: affixed form/base pairs
  - Window: 1 sentence

Wf processes explored

- "Baayen prefixes"
- Mean productivity rank assigned by 4 English morphologists:

un	1.500
re	1.625
mis	3.250
de	3.625
be	5.875
en	5.875
in	6.250

Extraction of prefixed forms

- From BNC

Extraction of prefixed forms

- From BNC
- All forms that begin with prefix string and with base that:

Extraction of prefixed forms

- From BNC
- All forms that begin with prefix string and with base that:
  - is at least 3 letters long

Extraction of prefixed forms

- From BNC
- All forms that begin with prefix string and with base that:
  - is at least 3 letters long
  - occurs in the corpus as independent word

Extraction of prefixed forms

- From BNC
- All forms that begin with prefix string and with base that:
    - is at least 3 letters long
    - occurs in the corpus as independent word
- E.g., *beads* is treated as prefixed; *bead* and *benny* are not

Extraction of prefixed forms

- From BNC
- All forms that begin with prefix string and with base that:
  - is at least 3 letters long
  - occurs in the corpus as independent word
- E.g., *beads* is treated as prefixed; *bead* and *benny* are not
- More "sophisticated" methods (that rely on further automated processing) perform worse than this "brutal" approach

Averaging Cosine/MI across forms with same prefix

- Cosine/MI computed for each prefixed form/base pair, but we want single value of each measure per prefix

Averaging Cosine/MI across forms with same prefix

- Cosine/MI computed for each prefixed form/base pair, but we want single value of each measure per prefix
- For each prefix class, compute average cosine/MI of 20 pairs with highest cosine/MI value

Averaging Cosine/MI across forms with same prefix

- Cosine/MI computed for each prefixed form/base pair, but we want single value of each measure per prefix
- For each prefix class, compute average cosine/MI of 20 pairs with highest cosine/MI value
- Rationale: presence of subset of prefixed words with high semantic transparency is more significant than fact that other forms in same class are opaque

Averaging Cosine/MI across forms with same prefix

- Cosine/MI computed for each prefixed form/base pair, but we want single value of each measure per prefix
- For each prefix class, compute average cosine/MI of 20 pairs with highest cosine/MI value
- Rationale: presence of subset of prefixed words with high semantic transparency is more significant than fact that other forms in same class are opaque
- E.g., *re-* has plenty of both transparent and opaque forms

Averaging Cosine/MI across forms with same prefix

- Cosine/MI computed for each prefixed form/base pair, but we want single value of each measure per prefix
- For each prefix class, compute average cosine/MI of 20 pairs with highest cosine/MI value
- Rationale: presence of subset of prefixed words with high semantic transparency is more significant than fact that other forms in same class are opaque
- E.g., *re-* has plenty of both transparent and opaque forms
- NB: hapax legomena are not playing a (crucial) role!

Results

- Morphologist's rank:
  - un, re
  - mis, de
  - be, en, in

Results

- Morphologist's rank:
  - un, re
  - mis, de
  - be, en, in
- $\mathscr{P}$:
  un $>$ re $>$ de $>$ mis, in $>$ en $>$ be

Results

- Morphologist's rank:
  - un, re
  - mis, de
  - be, en, in
- $\mathscr{P}$:
  un $>$ re $>$ de $>$ mis, in $>$ en $>$ be
- Cosine:
  un $>$ re $>$ in $>$ de $>$ mis $>$ en $>$ be

Results

- Morphologist's rank:
  - un, re
  - mis, de
  - be, en, in
- $\mathscr{P}$:
  un $>$ re $>$ de $>$ mis, in $>$ en $>$ be
- Cosine:
  un $>$ re $>$ in $>$ de $>$ mis $>$ en $>$ be
- MI:
  un $>$ re $>$ in $>$ de $>$ en $>$ mis $>$ be

Discussion

- Good results, especially with cosine similarity

Discussion

- Good results, especially with cosine similarity
- However, small data-set

Discussion

- Good results, especially with cosine similarity
- However, small data-set
- More nuance needed: polysemy of *in-*, *de-* vs. *deXize*

Current work

On Italian *ri-*

- Inspect corpus contexts to determine distribution and scope of senses (iterative vs. restitutive)

Current work
On Italian *ri-*

- Inspect corpus contexts to determine distribution and scope of senses (iterative vs. restitutive)
- Measure co-occurrence in relational patterns determined by mini-grammar (e.g., `V ART? ADJ* N`)

Current work

On Italian *ri-*

- Targets:
  - Single prefixed words
  - Class of *ri-* words (compared to other prefixed/non-prefixed words)
  - Iterative vs. restitutive

Current work
On Italian *ri-*

- Targets:
  - Single prefixed words
  - Class of *ri-* words (compared to other prefixed/non-prefixed words)
  - Iterative vs. restitutive
- Patterns of co-occurrence (both similarities and differences):
  - With bases (or prefixed forms with same bases)
  - With other *ri-* forms
  - With base+*again*
  - Direct co-occurrence with words that tap into the semantics of *ri-*

An encouraging pilot study

- Hypothesis: words containing more transparent/productive prefixes will have higher semantic/distributional similarity to other words containing same prefix

An encouraging pilot study

- Hypothesis: words containing more transparent/productive prefixes will have higher semantic/distributional similarity to other words containing same prefix
- *ri-* is more productive than *de-* in Italian (excluding *deXizzare* pattern)

An encouraging pilot study

- Hypothesis: words containing more transparent/productive prefixes will have higher semantic/distributional similarity to other words containing same prefix
- *ri-* is more productive than *de-* in Italian (excluding *deXizzare* pattern)
- Prediction: on average, *ri-* words will have more *ri-* words in their distributionally defined nearest neighbor set than *de-* words will have other *de-* words

Distributional data

- 1.9B token Web-crawled itWaC corpus (Baroni and Ueyama 2006)

Distributional data

- 1.9B token Web-crawled itWaC corpus (Baroni and Ueyama 2006)
- Automated thesaurus function of Word Sketch Engine (Kilgarriff et al. 2004)

Distributional data

- 1.9B token Web-crawled itWaC corpus (Baroni and Ueyama 2006)
- Automated thesaurus function of Word Sketch Engine (Kilgarriff et al. 2004)
- Based on Lin's (1998) distributional similarity measure

Lin's algorithm

- Collect collocates of each target word with other words in small set of grammatically meaningful patterns (e.g., for V collects N collocates in patterns `N ADJ* ADV* AUX* V`, `V ART* ADV* ADJ* N`, etc.)

Lin's algorithm

- Collect collocates of each target word with other words in small set of grammatically meaningful patterns (e.g., for V collects N collocates in patterns `N ADJ* ADV* AUX* V`, `V ART* ADV* ADJ* N`, etc.)
- For each pair of target words (with same POS), compute score based on number of shared collocates, weighted by MI (so that more unusual collocates will have more weight)

Lin's algorithm

- Collect collocates of each target word with other words in small set of grammatically meaningful patterns (e.g., for V collects N collocates in patterns `N ADJ* ADV* AUX* V`, `V ART* ADV* ADJ* N`, etc.)
- For each pair of target words (with same POS), compute score based on number of shared collocates, weighted by MI (so that more unusual collocates will have more weight)
- Pick as neighbor set of a target word all other target words with similarity score above a certain threshold (I used WSE defaults)

Neighbor sets

- Neighbor sets built in this way for our test words range from 31 to 59 members

Neighbor sets

- Neighbor sets built in this way for our test words range from 31 to 59 members
- E.g., neighbor set of *ricomporre* ("to recompose") include *ricostituire* ("to reconstitute"), *scomporre* ("to decompose"), *assemblare* ("to assemble")

Experiment

- 10 prefixed words with *ri-* and *de-* (but not *deXizzare*) randomly chosen from corpus (conditions: min fq $\geq$ 500, not in top 500 forms with prefix)

Experiment

- 10 prefixed words with *ri-* and *de-* (but not *deXizzare*) randomly chosen from corpus (conditions: min fq $\geq$ 500, not in top 500 forms with prefix)
- Number of forms with same prefix in neighbor sets:

	min	med	mean	max
ri-	3	14	13.4	25
de-	2	3	3.4	6

Experiment

- 10 prefixed words with *ri-* and *de-* (but not *deXizzare*) randomly chosen from corpus (conditions: min fq $\geq$ 500, not in top 500 forms with prefix)
- Number of forms with same prefix in neighbor sets:

	min	med	mean	max
ri-	3	14	13.4	25
de-	2	3	3.4	6

- Percentage of forms with same prefix over total number of neighbors:

	min	med	mean	max
ri-	10%	31%	28%	44%
de-	4%	8%	8%	13%

"Need" in the interpretation of productivity

- Ongoing work with Anke Lüdeling and Stefan Evert

“Need” in the interpretation of productivity

- Ongoing work with Anke Lüdeling and Stefan Evert
- Corpus counts are influenced by the need to express a given thought/concept

"Need" in the interpretation of productivity

- Ongoing work with Anke Lüdeling and Stefan Evert
- Corpus counts are influenced by the need to express a given thought/concept
- *Words are only formed as and when there is a need for them [. . .]* (Bauer 2001, 143)

"Need" in the interpretation of productivity

- Ongoing work with Anke Lüdeling and Stefan Evert
- Corpus counts are influenced by the need to express a given thought/concept
- *Words are only formed as and when there is a need for them [...]* (Bauer 2001, 143)
- The need to express something depends on extra-linguistic factors (like the political situation, fashion, etc.)

“Need” in the interpretation of productivity

- Ongoing work with Anke Lüdeling and Stefan Evert
- Corpus counts are influenced by the need to express a given thought/concept
- *Words are only formed as and when there is a need for them [. . .]* (Bauer 2001, 143)
- The need to express something depends on extra-linguistic factors (like the political situation, fashion, etc.)
- There is no *baayenitis* without Baayen!

Competition

- If a wf process is the *only* way to express a need, $\mathscr{P}$ will to a large extent measure extra-linguistic need, not factors relating to linguistic productivity

Competition

- If a wf process is the *only* way to express a need, $\mathscr{P}$ will to a large extent measure extra-linguistic need, not factors relating to linguistic productivity
- Need can be productive for extra-linguistic reasons!

Competition

- If a wf process is the *only* way to express a need, $\mathscr{P}$ will to a large extent measure extra-linguistic need, not factors relating to linguistic productivity
- Need can be productive for extra-linguistic reasons!
- The study of productivity is interesting only when studying relative close (interchangeable) competitors expressing the same need, as need factor is kept constant

Competition

- If a wf process is the *only* way to express a need, $\mathscr{P}$ will to a large extent measure extra-linguistic need, not factors relating to linguistic productivity
- Need can be productive for extra-linguistic reasons!
- The study of productivity is interesting only when studying relative close (interchangeable) competitors expressing the same need, as need factor is kept constant
- However. . . does competition exist? (cf. Plag 1999: where all have the rivals gone?)

Competition

- If a wf process is the *only* way to express a need, $\mathscr{P}$ will to a large extent measure extra-linguistic need, not factors relating to linguistic productivity
- Need can be productive for extra-linguistic reasons!
- The study of productivity is interesting only when studying relative close (interchangeable) competitors expressing the same need, as need factor is kept constant
- However. . . does competition exist? (cf. Plag 1999: where all have the rivals gone?)
- Probably it does (ongoing work on a set of German compound heads meaning "too much" with a disease connotation)

Outline

Conclusion

- The work of Baayen and others on productivity of fundamental historical importance:

Conclusion

- The work of Baayen and others on productivity of fundamental historical importance:
  - Early corpus-based work

Conclusion

- The work of Baayen and others on productivity of fundamental historical importance:
  - Early corpus-based work
  - Theoretical relevance of quantitative data

Conclusion

- The work of Baayen and others on productivity of fundamental historical importance:
  - Early corpus-based work
  - Theoretical relevance of quantitative data
  - Methodological expertise in using corpus data to answer linguistic questions

Conclusion

- The work of Baayen and others on productivity of fundamental historical importance:
  - Early corpus-based work
  - Theoretical relevance of quantitative data
  - Methodological expertise in using corpus data to answer linguistic questions
  - Development of descriptive techniques (VGCs etc.) and index ($\mathscr{P}$) to explore productivity in data-set

Conclusion

- Corpus-based productivity measures, qualitative explanations often *not* based on corpus evidence

Conclusion

- Corpus-based productivity measures, qualitative explanations often *not* based on corpus evidence
- Productivity measures have played and can play an important role in choosing productive phenomena to focus on (Baayen and Lieber 1991, Plag 1999, Lieber 2004)

Conclusion

- Corpus-based productivity measures, qualitative explanations often *not* based on corpus evidence
- Productivity measures have played and can play an important role in choosing productive phenomena to focus on (Baayen and Lieber 1991, Plag 1999, Lieber 2004)
- Possible criterion also in preparation of L2 teaching/lexicographic materials

Conclusion

- Corpus-based productivity measures, qualitative explanations often *not* based on corpus evidence
- Productivity measures have played and can play an important role in choosing productive phenomena to focus on (Baayen and Lieber 1991, Plag 1999, Lieber 2004)
- Possible criterion also in preparation of L2 teaching/lexicographic materials
- Many other areas of application still to explore: e.g., non-morphological productivity in studies of lexical richness, stylometry

Conclusion

- Very little corpus-based work on explaining productivity

Conclusion

- Very little corpus-based work on explaining productivity
- Corpora used as word frequency lists, context not taken into account

Conclusion

- Very little corpus-based work on explaining productivity
- Corpora used as word frequency lists, context not taken into account
- Typically, coarse level of analysis (e.g., prefix polysemy ignored), probably in part due to data sparseness, in part to manual work demands

Conclusion

- Very little corpus-based work on explaining productivity
- Corpora used as word frequency lists, context not taken into account
- Typically, coarse level of analysis (e.g., prefix polysemy ignored), probably in part due to data sparseness, in part to manual work demands
- Reasons to be optimistic:

Conclusion

- Very little corpus-based work on explaining productivity
- Corpora used as word frequency lists, context not taken into account
- Typically, coarse level of analysis (e.g., prefix polysemy ignored), probably in part due to data sparseness, in part to manual work demands
- Reasons to be optimistic:
  - Availability of very large corpora

Conclusion

- Very little corpus-based work on explaining productivity
- Corpora used as word frequency lists, context not taken into account
- Typically, coarse level of analysis (e.g., prefix polysemy ignored), probably in part due to data sparseness, in part to manual work demands
- Reasons to be optimistic:
  - Availability of very large corpora
  - Automated corpus-based grammatical/semantic analysis methods from NLP

Conclusion

- Very little corpus-based work on explaining productivity
- Corpora used as word frequency lists, context not taken into account
- Typically, coarse level of analysis (e.g., prefix polysemy ignored), probably in part due to data sparseness, in part to manual work demands
- Reasons to be optimistic:
  - Availability of very large corpora
  - Automated corpus-based grammatical/semantic analysis methods from NLP
  - New analytical tools to study context and meaning from corpus linguistics, e.g., Stefanowitsch and Gries' (2005 and elsewhere) collustructional analysis

THE END