Some views on, and suggestions for, quantitative corpus linguistics

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

Introduction: profile and goals

Corpora that I work with, and what I do with them

Methodology: more & better regressions

Goals that I have for the field as a whole

Theory: go beyond frequency

Concluding remarks

My corpus-linguistic profile

- The corpora I work with most
  - synchronic native-speaker data (BNC, ICE-GB, [BF]rown, P(EB)ob.)
  - synchronic learner corpora (ICLE, LOCNESS, LINDSEI, …)
  - synchronic indigenized-variety data (ICE-…, SAVE, …)
  - sometimes diachronic/longitudinal data (Helsinki corpus, PCEEC, …)
  - sometimes learner corpora (CHILDES)
  - corpus methods I often employ
    - concordance data
    - sometimes frequency lists
    - sometimes dispersion
  - sometimes diachronic/longitudinal data (Helsinki corpus, PCEEC, …)
  - sometimes learner corpora (CHILDES)
  - corpus methods I often employ
    - concordance data
    - sometimes frequency lists
    - sometimes dispersion
  - statistical methods I often employ
    - all sorts of regression modeling
      - fixed-effects, mixed-effects, (generalized) (linear|additive) modeling
      - other classifiers (CART, Random Forests, …)
      - (clustering, PCA, MDS, …)

One reason we need regressions ...

Let's assume one is interested in how media coverage of the word "Muslim" changes over time (any resemblance to real studies, published on-going, is not coincidental)

Let's assume this is backed up by a correlation measure: r=0.97, p<0.001

However, one needs the interaction WORD:TIME ...

Some views, and suggestions for, quantitative corpus linguistics

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

What are your goals?

- My main goal is to change corpus linguistics
  - by making it less humanities and more social science
  - no, corpus linguistics is not just looking at the text
  - no, corpus linguistics is not just some hermeneutic circle
  - yes, corpus linguists may want to generalize
  - yes, corpus linguistics may want to connect to more social sciences
  - yes, corpus linguistics needs statistics & experimental work
  - by making it realize more of its connections to
    - theoretical linguistics (esp. cognitive/usage-based ling.)
    - psycholinguistics
  - by making it realize more of its potential to connect with, and learn from, other disciplines (within limits)
    - computational linguistics
    - digital humanities-related research
    - ecology,
  - by making it methodologically better and more aware by recognizing what happens in neighboring fields
    - theoretically (see above)
    - methodologically and in particular statistically
    - in particular statistically

Some views, and suggestions for, quantitative corpus linguistics

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

Introduction: profile and goals

Methodology: more & better regressions & experimental work

Goals that I have for the field as a whole

Theory: go beyond frequency

Concluding remarks

One reason we need regressions ...

Let's assume one is interested in how media coverage of the word "Muslim" changes over time (any resemblance to real studies, published on-going, is not coincidental)

Let's assume this is backed up by a correlation measure: r=0.97, p<0.001

However, one needs the interaction WORD:TIME ...

Some views, and suggestions for, quantitative corpus linguistics

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries

One reason we need regressions ...

Let's assume one is interested in how media coverage of the word "Muslim" changes over time (any resemblance to real studies, published on-going, is not coincidental)

Let's assume this is backed up by a correlation measure: r=0.97, p<0.001

However, one needs the interaction WORD:TIME ...

Some views, and suggestions for, quantitative corpus linguistics

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
http://tinyurl.com/stgries
Another reason we need regressions ...

- However, regressions are also useful for seemingly much simpler applications where ppl usually use \(X^2\).
  \(X^2\) is not a good thing for two reasons:
  - simplification (not just as above) and missing opportunities
  - independence of data points

- imagine you study genitives in varieties of English and rights-take

**approach 1:** \(X^2\)
  - \(X^2=11.01, df=2, p<0.005\)
  - \(G=11.01, \chi^2=0.04\)
  - this misses out on a lot

- contrasts, confidence intervals, straightforward visualization comparable to other studies

**approach 2:** glm
  - with \(X^2\) treatment contrasts
    - \(p<0.006\)
  - with orthogonal contrasts
    - nat vs indig: \(p<0.0014\)
    - \(\text{coef}=0.19, p<0.0015\)
    - \(\text{coef}=0.13, p<0.0018\)
    - \(\text{coef}=0.05, p=0.42\)
  - \(R^2, Cc=0.52\)

- And if you decide to do regressions, do them the right way:

**mixed-effects models** have become more widely used but analyses routinely leave things to be desired:
- often, categorical predictors are explored with only straight-line effects (rather than curved ones)
- example: let's try to predict whether NNSs of Japanese realize
  the subject of clauses like NSs do
- predictors:
  - \text{GIVENNESS (0:10) of subject}
  - \text{CONTRAST (n/y) of subject}
  - \text{optimal random-effects structure}
  - one result: the more the NNSs are nativelike
  - let's analyze this again, but now with
  - poly(GIVENNESS (0:10) of subject, 2)
  - \text{CONTRAST (n/y) of subject}
  - same random-effects structure
  - result: the more subjective givenness

**Theoretically, we also need to**

cast a wider net ...

- Consider, for example, how frequency has assumed a central role in cognitive/usage-based linguistics
  - it has been assumed for a long time now that token frequency (co-)determines degree of entrenchment ...
  - which (co-)determines speed/ease of lexical access, retrieval, production ...

- "Continuous scale of entrenchment in cognitive organization. Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence." (Moran 1987:59)

- "[T]his seems highly convincing, not least in view of the considerable body of evidence from psycholinguistic experiments suggesting that frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency and frequency co-varies." (Schmid 2010a:115f.)
Kinds of frequencies

- We can distinguish different kinds of frequencies
  - Conceptual frequency (see Hoffmann 2004)
  - Type frequency
  - Token frequency

Following Schmid, token frequency then can be divided into

- Absolute frequency (→ context-free entrenchment)
  - Counts of x in a (maybe normalized) corpus.

- Relative frequency (→ contextual entrenchment)
  - Counts of x with context y in a corpus

For example,

- Abs. freq. of x: a + b
  - Rel. freq. of x given y: \( \frac{a}{a+b} \)

It seems as if most of usage-based linguistics focuses on the latter

\[ p(\text{unit } x | \text{unit } y) \]

(Schmid discusses reasons to also consider co-text freq)

On the other hand, maybe token frequency is less important than usually thought

Ellis (2011) has emphasized that

"it is contingency, not temporal pairing, that generated conditioned responding" in classical conditioning

"human learning is [...] perfectly calibrated with normative stats. measures of contingency like r, \( \chi^2 \) & \( \Delta \rho \)

\[ p(\text{unit } x | \text{unit } y) \]

"[l]anguage learning can thus be viewed as a statistical process in that it requires the learner to acquire a set of likelihood-weighted associations between constructions and their functional/semantic interpretations"

- Shillcock & McDonald’s (2001) CD outperforms abs. freq. as a predictor of lexical decision times

Baayen (2010:436)

- The word frequency effect in the sense of pure repeated exposure accounts for only a small proportion of the variance in lexical decision

- Local syntactic & morphological co-occurrence probabilities are what makes word frequency a powerful predictor for lexical decision latencies

On contingency...

The following are considerations that are relevant to choosing a measure of contingency/association

- Symmetry: is the AM supposed to be symmetric or not?
  - Nearly all AMs are: \( p_\text{rel} \), \( LLR \), \( X^2 \), \( MI \), \( t \), \( z \), odds ratio, ...

- Some are not: \( p(y|x) \), \( p(x|y) \), \( \Delta p \)

- Metric type: +effect – freq. vs +effect +frequency
  - The former: odds ratio, the asymmetric ones from above, ...

- The latter: \( LLR \), \( X^2 \), ...

- Frequency information: token vs token-type frequency
  - The former: all but one

- The latter: lexical gravity, \( G \)

- Dispersion: is dispersion information included? so far, virtually never probably best settings in an ideal world:

  - Symmetry: no
  - Metric type: +effect
  - Frequency: token-type
  - Dispersion: included (see below)
• Most corpus-based work explores context by studying/annotating concordances and then compute freg/perms. While this allows to see minute details, it may also lead to easy overestimates of the role of frequency, obscuring bigger distributional trends.

• McDonald & Shillcock (2001) suggest that what seems like frequency effects may in fact be epiphenomenal— they propose the measure Contextual Distinctiveness (CD), $D_C(P|Q) = D(P_{(to posterior)}||P_{(from prior)})$ for words, e.g., that is the (asymmetric!) divergence from the probability distributions of words in a corpus to the probability distributions of words in some context with some function. Crucially, this measure is correlated with frequ abs but does not enter into its computation.

• It accounts for variance in RTs when word freq & length are controlled for whereas freq does not when length and CD are controlled for.

> Some views, and suggestions for, quantitative corpus linguistics University of California, Santa Barbara

On recency/dispersion

- Let us return to Schmid (2010): “frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention” (Recency & Contextual Diversity, 2006) is important because of how it relates to acquisition, learning, and forgetting: “the extent to which the number of repeated exposures to a particular item affects that item’s later retrieval depends on the separation of the exposures in time and context” (ABQ – hardly ever utilized outside of the context of priming: either a better or at least an additional measure of how likely it is something is encountered by a speaker, e.g., the words enormous and staining are equally freq in the Brown corpus but as unequally dispersed as possible.

> BNC-spoken: council = same freq bin as staining are equally freq in the Brown corpus but as unequally dispersed as possible.

• Crucially, this measure is correlated with freq abs; does not enter into its computation. It accounts for variance in RTs when word freq & length are controlled for whereas freq does not when length and CD are controlled for.

- 1st, there are more domain-general mechanisms that require corpus-linguistic operationalizations.

- 2nd, corpus linguistics must look more to what other disciplines do with corpus data, e.g., cognitive and, more importantly, psycholinguistics.

- 3rd, cognitive linguistics must stop underutilizing corpus data – raw freqs & pros aren’t all there is. Let’s not fall into the 1990s polysemy-network trap and, now, say ‘everything’s due to frequency.’ This is not just number-crunching but also theoretically important.

- If frequency is a causal factor (as a repetition counter), then maybe entrenchment/resting-level accounts are correct.

- If it isn’t and it is contiguity, entropy, etc., then we probably need different psycholinguistic models.

- 4th, let’s drop the pessimism before we have explored more and better options.

Thank you!

http://tinyurl.com/stgries