An aspect of the competition between *-ity* and *-ness*

Kaarel Hänni

24.914 Spring 2021

Introduction

One way to assess the productivity of an affix over time is to look at the number of neologisms with that affix that appear at each year in a corpus. In particular, we can use this neologism metric to get an understanding of the productivity of *-ity* as compared to the productivity of *-ness*. To understand the competition between these two affixes, another thing one can look at are particular pairs of words obtained from the same stem by affixing *-ity* or *-ness* (in cases where both words are used a decent amount), tracking the frequencies over time. An example of such a pair would be *fashionability* and *fashionableness*. One might guess that there would be a correlation between the relative trends in the two metrics. That is, for time periods where the neologism metric suggests that *-ity* is doing relatively well against *-ness*, we might also expect a particular *-ity* word to be doing relatively well against the *-ness* word with the same stem. The aim of this paper is to make some progress towards finding out if this is true.[1]

To address this question, I will be looking at data from the Google Ngrams corpus. One caveat to note here is that the Ngrams corpus only has information about words that appeared in at least 40 books, meaning that we will only be able to find neologisms that were later used in at least 39 other books in the corpus as well. The only data I will be analyzing is in the files `ity-eng-us-all-1gram-20120701.filtered.txt` and `ness-eng-us-all-1gram-20120701.filtered.txt` from Canvas. I am going to analyze data from these two files to generate for each year i a parameter $x_i \in [0, 1]$ that describes how often *-ity* is seen in new words vs how often *-ness* is seen in new words, as well as a parameter $y_i \in [0, 1]$ that describes how often the *-ity* word is used vs how often the *-ness* word is used for same-stemmed pairs where both words have a relatively high frequency. Explicit definitions of these parameters will be given in later sections. Once I have x_i and y_i for each year, I will see if there is a correlation.

[1] This was also one of the questions suggested in the prompt.

How the *-ity* vs *-ness* neologism productivity parameters are calculated

I will now explain what I am defining the x_i parameters to be. For each year i, we want x_i to be a real number between 0 and 1 that is small when there are few *-ity* neologisms compared to the number of *-ness* neologisms, and large when there are many *-ity* neologisms compared to the number of *-ness* neologisms. A simple way to accomplish this is to compute the following:

- For each year i, find out how many words in `ity-eng-us-all-1gram-20120701.filtered.txt` appeared for the first time in year i; let this number be $Q_{ity,i}$.
- Compute $Q_{ness,i}$ analogously. That is, for each year i, find out how many words in `ness-eng-us-all-1gram-20120701.filtered.txt` appeared for the first time in year i; let this number be $Q_{ness,i}$.
- Compute
$$x_i := \frac{Q_{ity,i}}{Q_{ity,i} + Q_{ness,i}}$$
In words, x_i is the fraction of all new *-ness* or *-ity* words (in year i) that are *-ity* words.

A concern here is that a word appearing for the first time in the corpus does not necessarily imply that the writer formed the word themselves using their grammar. They might still be remembering the word (from some source not included in the corpus). To address this issue, I will be focusing on years after 1850. The corpus has a good amount of data in total from years before 1850, and I hope that this limits the number of words appearing for the first time in the corpus but not being recent innovations. (At least from visual inspection, no large drop is noticeable at the start of this period, although this is quite poor evidence.) Additionally, there is some hope that when computing x_i, there would be cancellation between such non-innovative corpus neologisms' contributions to $Q_{ity,i}$ and $Q_{ness,i}$. I will set the end of the time period I will look at to be 1980. This avoids divisions by zero, as well as generally small numbers of neologisms.

I carried out this computation (and all other data processing used in this paper) in Python. See Figure 1 for a plot of x_i from 1850 to 1980. Interestingly, from this graph, it looks like *-ity* is becoming more productive than *-ness* over time. (This is confirmed by a linear fit with slope 0.00241, $r = 0.593$ and p-value 10^{-13}, assuming I am doing the statistics correctly, which I am not entirely sure about, since I have a suspicion that the p-value here is computed without taking into account some correlations between residuals which are often important for time-series. But I also have a suspicion that a better way of doing this would still give a tiny p-value.) In fact, more than half of the neologisms captured by this analysis appear to be *-ity* forms. This contrasts with the findings of Baayen and Renouf, who found that *-ness* is more productive that *-ity*, and increasing in relative productivity. One way to reconcile these two facts is to say that

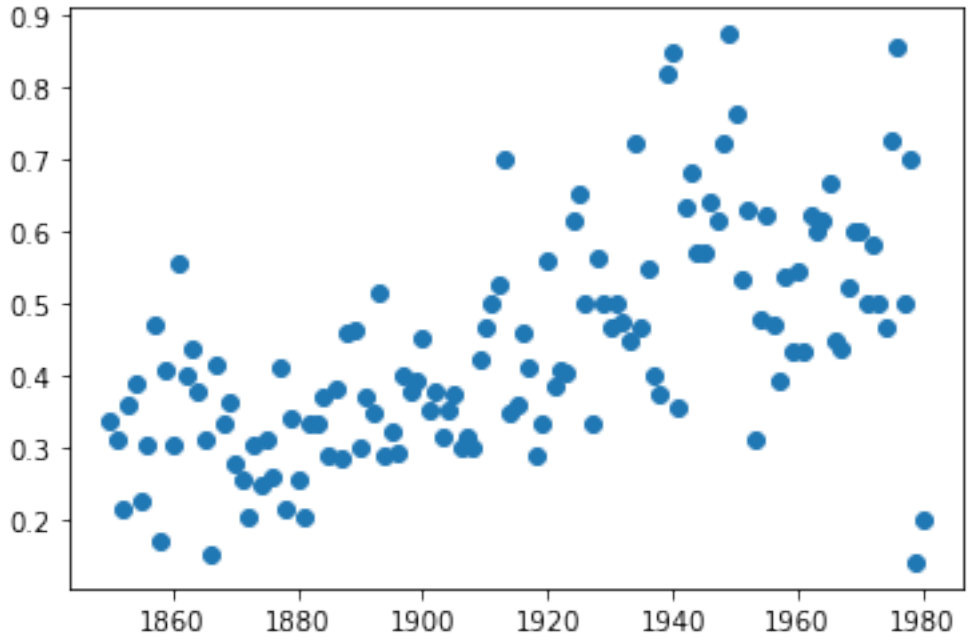

Figure 1: Fraction of *-ity* neologisms by year

perhaps there are actually more *-ness* neologisms over time (and their relative frequency might be increasing), but these have a harder time making it above the threshold of mentions in 40 different books in the Google Ngrams corpus. In other words, it could be that there are fewer *-ity* neologisms, but the few that exist gain popularity quicker. Perhaps this might have something to do with the claim that each new *-ity* form is more often memorized upon encountering it, but each *-ness* form has to be independently generated a larger number of times to make the cut of 40. It seems to me entirely possible that this alone would explain the upward trend we are seeing, in which case the graph might not say much about the relative productivity of the two affixes. That said, from now on, I will ignore this (possibly crucial) problem until coming back to it briefly at the end of this paper.

How the frequent *-ity* vs *-ness* form-by-form competition parameters are calculated

I will now explain how I am definining the *-ity* vs *-ness* parameters that should contain information about same-stemmed pairs. I will start by explaining how I found the pairs (like *emotivity* and *emotiveness*) to look at. I started by restricting the full lists of *-ity* and *-ness* verbs in the corpus to only include

those words with total frequency at least 1000.[2] I then removed the last 3 letters from each *-ity* word and the last 4 letters from each *-ness* word, and then for each word in the first list of stems, I searched the second list for stems that are closer than 1 Levenshtein distance to it. As I wanted to check all pairs by inspection at the end and the number was still a bit unwieldy, I then ran the remaining pairs through a few other filters. Namely, I removed all stems of length 2 or less and also pairs for which the first letter did not match. To keep only pairs in which interesting variation between the two forms is likely to be seen, I only kept those where the total frequencies of the two forms in the corpus do not differ by more than a factor of 5. After manually getting rid of a number of remaining nonsense pairs (e.g. coming from likely unintentional spellings, such as *availablity* (without the *i*) and *availableness*, or pairs with likely different meanings, e.g. *bumpity* and *bumpiness*), I was left with 36 pairs. The list is given in the appendix.[3]

I will now define the y_i parameters that are supposed to track the competition between the two forms of these pairs.

- Let the pairs be indexed by $j = 1, 2, \ldots, 36$.
- Let $Q_{ity,j,i}$ be the number of pages the *-ity* word in pair j appears on in year i, as indicated in `ity-eng-us-all-1gram-20120701.filtered.txt`.[4]
- Let $Q_{ness,j,i}$ be the number of pages the *-ness* word in pair j appears on in year i, as indicated in `ness-eng-us-all-1gram-20120701.filtered.txt`.
- For each $j = 1, \ldots, 36$ for which $Q_{ity,j,i} + Q_{ness,j,i} \neq 0$, let

  $$y_{j,i} = \frac{Q_{ity,j,i}}{Q_{ity,j,i} + Q_{ness,j,i}}.$$

  In words, $y_{j,i}$ is the fraction of all words from the pair used in year i that are the *ity* word.
- Finally, let y_i be the average of $y_{j,i}$ over all m pairs j for which $y_{j,i}$ is defined for year i. That is,

  $$y_i = \frac{\sum_{j=1}^{36} y_{j,i}}{m}.$$

  (Fortunately, $m = 0$ does not happen for any year i.)

See Figure 2 for a plot of y_i from 1850 to 1980. Interestingly, we again note that there seems to be a general upward trend. It seems to be even clearer this time.) (A linear regression with slope 0.00205, $r = 0.867$ and p-value 10^{-40} confirms this, with the same caveat as for the last slope calculation.)

[2] The numbers in this paragraph are chosen by trial and error so as to end up with a sample of pairs of manageable size.

[3] I could have instead used the summary file for finding the stems, but the approach here was also interesting.

[4] I might instead have used the book count here, but that column of the files seemed to be 0 for some reason. Maybe I was confusing something.

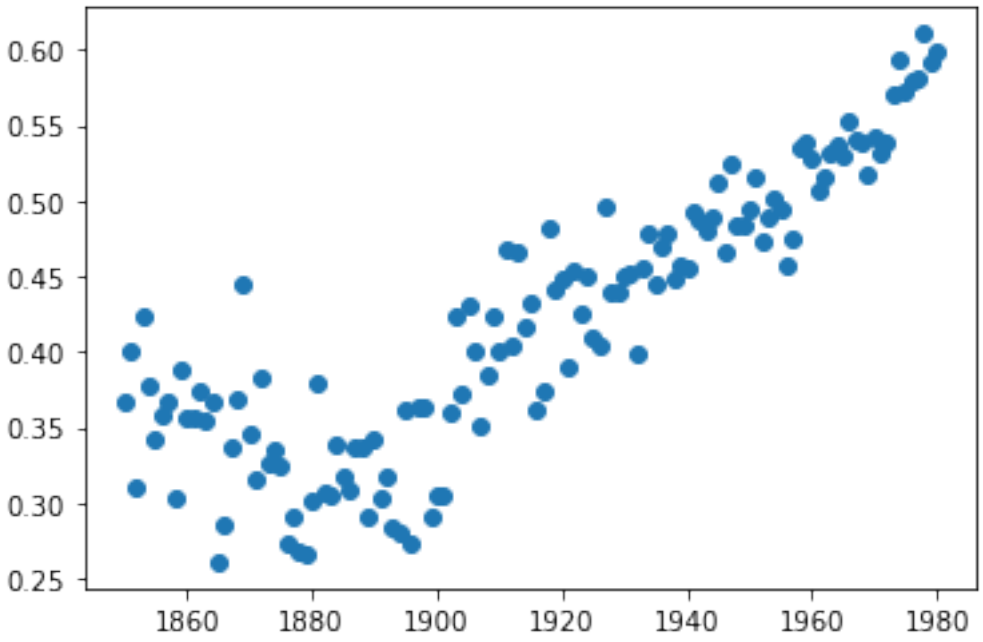

Figure 2: Form-by-form fraction of *-ity* words per year

Conclusions and further discussion

Now that we have x_i and y_i for all years from 1850 to 1980, it remains to just find out if x_i and y_i are correlated. See Figure 3 for a plot of x_i against y_i. The OLS linear fit between x_i and y_i has slope 0.323, $r = 0.555$, and p-value 10^{-11}. However, I think that this p-value is not entirely correct, because we are dealing with time-series data, and I would expect there to be correlations between adjacent residuals, which would change the standard deviation estimate that goes into the p-value calculation. Maybe Newey-West would be better, but as I would not expect this to change the p-value drastically and I am also not sure how to do this within a reasonable amount of time in Python, I will skip that.

A question that seems interesting to me is whether the regression coefficient on x_i remains nonzero after putting time in the controls. I think that according to the regression anatomy theorem, instead of regressing y_i on x_i and i simultaneously and finding if the coefficient on x_i is significantly different from 0, it is roughly equivalent to first regress x_i on i, compute the residuals $\tilde{x}_i = x_i - (a + bi)$, and then find if the coefficient in the regression of y_i on $\tilde{x}_i$ is significantly different from 0. (Anyway, hopefully this is close to the right thing, as multiple independent variable regressions currently seem to be a bit out of reach of my Python statistics skills.)

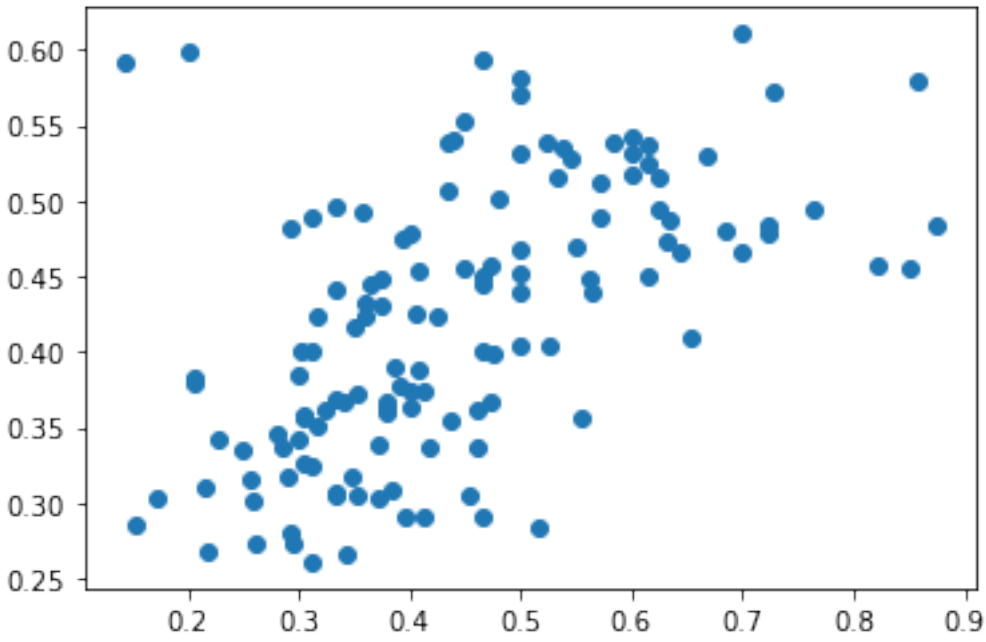

Figure 3: x_i on the horizontal axis, y_i on the vertical axis

Figure 4 shows the plot of $(\tilde{x}_i, y_i)$. The linear regression of y_i on these residuals has $p = 0.563$. So it seems that after controlling for a linear trend, we do not observe a statistically significant correlation between x_i and y_i. This is not what I expected.

I made a number of arbitrary choices in this analysis. It would be interesting to return to this question and to look at the Ngrams corpus more thoroughly, as well as at some other data sets. More data might decrease the standard errors. If there is a correlation even after controlling for time, then smaller standard errors might make that effect statistically significant in such an analysis.

Finally, it would be great to repeat the same analysis with hapaxes instead of neologisms. However, it can again be somewhat problematic that the Google Ngrams corpus only has words that have appeared in at least 40 books. Treating words that appear only a bit more than 40 times as hapaxes, it seems to me that one can still run into similar problems as with neologisms. It might be good to look at another corpus that includes lower-frequency words.

Appendix – competing pairs

Here is the list of pairs I used for calculating y_i:

- `[['africanity', 'africanness'],['coercivity', 'coerciveness'],['compulsivity',`

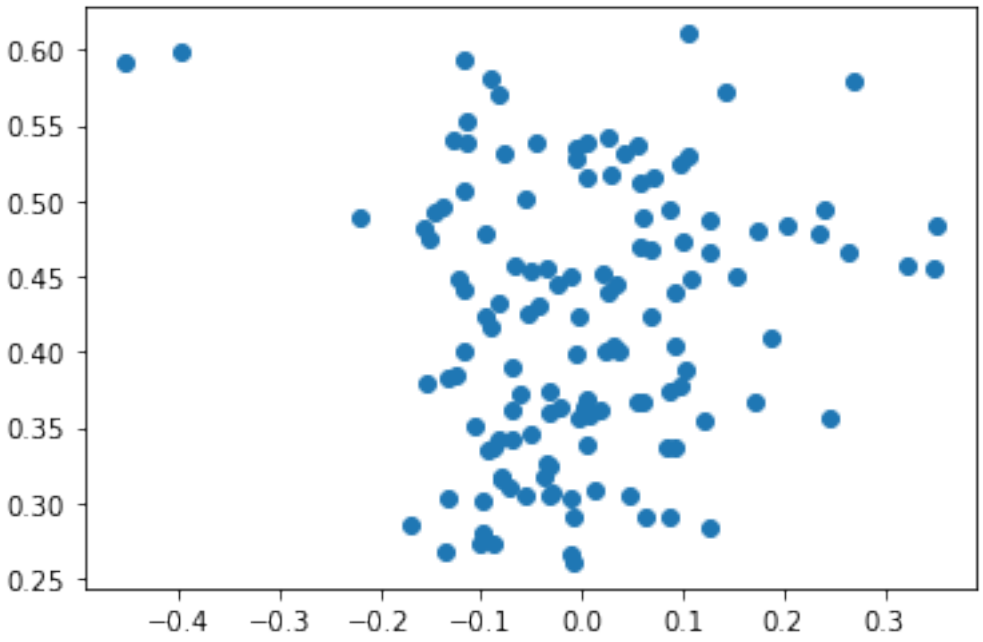

Figure 4: $\tilde{x}_i$ on the horizontal axis, y_i on the vertical axis

```
'compulsiveness'],['corrosivity','corrosiveness'], ['crudity','crudeness'],['curiousity
'curiousness'],['directivity', 'directiveness'],['emotivity',
'emotiveness'],['equivocality', 'equivocalness'],['exclusivity',
'exclusiveness'],['expressivity', 'expressiveness'],['floridity',
'floridness'],['impassivity', 'impassiveness'],['impulsivity',
'impulsiveness'],['ineffectuality', 'ineffectualness'],['liablity',
'liableness'],['mundanity', 'mundaneness'],['nebulosity', 'nebulousness'],['numerosity'
'numerousness'],['operativity', 'operativeness'],['perceptivity',
'perceptiveness'],['perversity', 'perverseness'],['ponderosity',
'ponderousness'],['progressivity', 'progressiveness'],['radicality',
'radicalness'],['receptivity', 'receptiveness'],['reflectivity',
'reflectiveness'],['regressivity', 'regressiveness'],['reliablity',
'reliableness'],['religiosity', 'religiousness'],['reproductivity',
'reproductiveness'],['sanguinity', 'sanguineness'],['scrupulosity',
'scrupulousness'],['sensitivity', 'sensitiveness'],['tensity',
'tenseness'],['tepidity', 'tepidness']]
```