



## CogniTExtes

Revue de l'Association française de linguistique cognitive

Volume 19 | 2019

Corpora and Representativeness

---

# The importance of sampling frames in representative historical corpora : a case study of Parisian theater

Angus B. Grieve-Smith

---



### Electronic version

URL: <http://journals.openedition.org/cognitextes/1671>

DOI: 10.4000/cognitextes.1671

ISSN: 1958-5322

### Publisher

Association française de linguistique cognitive

### Electronic reference

Angus B. Grieve-Smith, « The importance of sampling frames in representative historical corpora : a case study of Parisian theater », *CogniTExtes* [Online], Volume 19 | 2019, Online since 17 June 2019, connection on 19 June 2019. URL : <http://journals.openedition.org/cognitextes/1671> ; DOI : 10.4000/cognitextes.1671

---

This text was automatically generated on 19 June 2019.



*CogniTExtes* est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

---

# The importance of sampling frames in representative historical corpora : a case study of Parisian theater

Angus B. Grieve-Smith

---

## 1. Introduction

- 1 Corpora are tools that we use to answer research questions about a particular aspect of language. When we ask whether a corpus is representative, we are asking whether observations about that corpus can be generalized to the variety of language that interests us. The key to representativeness is the sampling frame. Is the corpus sampled from the variety that concerns our research question? Or is the sampling frame of the corpus at least connected to the variety that interests us in some way that allows us to interpret conclusions about the sampling frame as answers to our research questions?
- 2 Theories of cognitive linguistics, particularly usage-based theories of language change, rest on the hypothesis that the language perceived by a person affects the language that they produce in the future. In this study, following research I conducted for my dissertation study, the Spread of Change in French Negation (Grieve-Smith 2009), I will focus on the claim that the relative type frequencies of constructions in competition in a language user's input predict changes in the relative type frequencies of these constructions in the output (Bybee 1995). To test this hypothesis, we need data from at least two periods: one set of data representing the output of language users, and another from a previous period representing the language that those same users experienced as input.
- 3 Testing these kinds of questions requires a corpus that is representative of both the input and output of particular language users, as described above. I will compare FRANTEXT (CNRTL 2018), the leading corpus of nineteenth-century French, with the Digital Parisian

Stage corpus, a one percent sample of the theatrical production of nineteenth-century Paris, based on Wicks (1950 et seq.). The very different results produced by these two corpora are not sufficient by themselves to test the type frequency hypothesis, but they have strong implications for the methods that we use to answer questions like this.

## 2. Sampling in population studies

- 4 Sampling in language corpora is built on a strong body of work in population studies, pioneered by Pierre-Simon Laplace. Laplace (1814: 45) noted that the French Empire did not have reliable population counts for the territory it controlled, but it did have reliable birth records for every municipality. He argued based on numerous cases that the ratio of total population to birth rate was relatively constant across Europe, so that if the government had an estimate of that ratio it could multiply that by the recorded birth rates to estimate the total population size.
- 5 Laplace then reasoned that the government could estimate the average birth rate by "choosing districts distributed in a roughly uniform manner throughout the Empire, in order to generalize the result, independently of local circumstances." He based his method on the Law of Large Numbers discovered by Jacob Bernoulli (1713), which asserts that the outcome of repeated executions of random events, such as spins of a roulette wheel, will tend to converge on the expected value of the events. Sampling the population size in a random district is exactly the kind of event that Bernoulli described (a Bernoulli trial), and measuring the population size (and then determining the ratio of population to birth rate in a sample with a large enough number of districts) would therefore yield an average that approached the average throughout the Empire.
- 6 In 1802, based on Laplace's recommendations, the then-French Republic conducted a population count of 30 out of its 108 districts, yielding an average ratio of 28.352843 inhabitants per annual birth. Laplace then used this ratio to estimate the total population of all 108 districts at 42,529,627, with a one in 1161 chance that the error was greater than half a million people (1.17% of the total). Laplace did not have a reliable census count to judge the validity of his estimate with, but his methods have been confirmed by years of follow-up research. They have also been applied beyond the field of population studies, including biology, medicine and agriculture (Student 1908: 22).
- 7 Laplace recommended this method because a full population census is "arduous and difficult to conduct accurately." This is an important observation to keep in mind in all discussions of sampling. Sampling is a labor-saving device, reducing the number of observations required to reach a conclusion with an acceptable level of accuracy.
- 8 The numerous tests of statistical significance developed by Student (1908) and others are all based on Laplace's sampling methods. They are refinements of Laplace's figure of one in 1,161: given a measurement and a sample size, what is the probability ( $p$ ) that the measurement is due to accidentally sampling unrepresentative members of the population? Is this probability within a range that is acceptable to us?

## 3. The sampling frame and the limits of generalization

- 9 There are several ways that Laplace's methods can be misapplied, yielding results that are inaccurate or misleading. The most problematic is to use sampling methods that are not

representative, and thus likely to incorporate bias in some way, such as a researcher sending a survey invitation to their Twitter followers. Another is to sample from a frame that is not applicable to the research question. This second misapplication is a particular challenge for corpus linguistics, and I will focus on it for the remainder of this paper.

- 10 In Laplace's estimate of the population of France in 1802, his research question was simple: "determine the population of a great empire." The sample was the 30 districts where the government took a census count, while the total 108 districts of the then-Republic constitute the sampling frame. The research question applied directly to the sampling frame, and the uniform sampling strategy justified generalizing the results from the sample to the entire frame.
- 11 Laplace's sampling frame, the 108 districts of the French Republic in 1802, was well-known, but in a study of language the appropriate sampling frame is not always so obvious. The Brown Corpus (Francis and Kučera 1964) was compiled on the basis of samples drawn from the Brown University Library catalog, meaning that the results of studies based on Brown can be generalized to the titles in that catalog.
- 12 Results from the Brown corpus can only be generalized to the English language as a whole to the extent that the Brown University Library catalog is itself representative of the English language. This is problematic, because the catalog is not a random sample of the English language. Each of those books was published or distributed based on a series of conscious decisions made by the publishers. The catalog is further filtered by the acquisition decisions of the University librarians. These two levels of bias preclude results based on the Brown corpus from being automatically generalized to the English language as a whole.
- 13 In my dissertation study of the spread of negation in French (Grieve-Smith 2009) I used a sample taken from FRANTEXT (Imbs 1971). This sample was a good start and the study indicated a promising direction of research, but as I studied the corpus and learned more about sampling methods, I came to the conclusion that it was not adequate to answer my research question. In the rest of this paper I will outline an alternative methodology that is more true to the example of Laplace's study, and more likely to provide satisfactory answers to our questions.

## 4. Research questions in cognitive linguistics

- 14 Laplace aimed to determine the population of the French Empire. In cognitive linguistics we have several research questions that we can test with corpus studies. One of our central principles is that the usage of language in one period is a factor in determining the usage of that language in future periods. In my dissertation, *The Spread of Change in French Negation* (Grieve-Smith 2009), I set out to test the hypothesis that the type frequency of a construction in one period determines the type frequency (and consequently the token frequency) in future periods.
- 15 The type frequency hypothesis rests on multiple observations (Bybee 1995: 451) that constructions that are used with a greater number of types tend to be associated with new types at a higher rate than other constructions that share the same function. The cognitive theory underlying it (Zager 1985: 136) is that a category with a greater diversity of members is perceived as being more open to new members. To my knowledge, this hypothesis had not previously been tested on a new corpus.

- 16 I combined Bybee and Zager's theory with Kroch's (1989) model of propagation in language change, and Lotka (1925) and Volterra's (1926) model of competition for resources. This led to one of the hypotheses for my Spread of Change study (Grieve-Smith 2009: 103): "When constructions compete for a function, the change in type frequency for each construction in a given period will be consistent with the Lotka-Volterra model based on that construction's type frequency in the preceding period."

## 5. Sampling frames for questions about language

- 17 In order to test a question like "Can changes in type frequency in one period be predicted by the type frequency in an earlier period?" we need to understand where and how to measure type frequency. We can measure the type frequency in the later period by measuring the output of language users during that period, but how do we measure the preceding period - the input to those language users?
- 18 A language user can derive their sense of relative type frequency or "openness to new members" from the language they perceive, or from the language they produce. It is possible that both experiences are relevant, and to my knowledge this question has not yet been fully tested. Bybee's theory rested on the input (Bybee 1995: 427), so this is what I chose to test for the Spread of Change study.
- 19 This leads us to the key challenge in the design of historical corpora, where the question of bias and sampling frames intersects with the challenge of adequately understanding language variation. Language varies according to multiple factors including region, social class, gender and situation. A corpus is likely to include the production of several individual language users, in multiple varieties, from every period. Each language user is exposed to multiple varieties over the course of their life.
- 20 How well do the language varieties present in the corpus at one period represent the varieties that the authors of texts present in a subsequent period were exposed to? There have been efforts to record the proportions of various varieties in the input of particular users, but it is obvious that historical corpora cannot reproduce these proportions well. One particular difficulty is the way that the genre balance of recorded texts changes over time. There were no tweets, IRC chat messages or telegrams before 1800, and relatively few epic poems and mystery plays after 1700.
- 21 Another difficulty is that every language user is exposed to multiple informal conversational language varieties; the bulk of language input for most speakers is likely in informal conversation. It may be possible for language users to compartmentalize the inputs from different varieties, but only to a limited extent; the history of language is full of examples of unconscious transfers of features from one variety to another. For any period before the widespread availability and affordability of voice recording, a corpus is highly unlikely to have these varieties represented in any proportion comparable to the input received by a typical language user.
- 22 There are strategies that we can use to compensate for the lack of informal conversation in our sampling frames. Some other varieties share features with informal conversation, either by accident or by design. One such variety is the language of theater. Many theatrical scenes are intended as reasonable facsimiles of informal conversations.
- 23 This is not to say that theatrical scripts can be relied on as faithful reproductions of informal conversational language (Lodge, 1991). They are usually planned in advance, for

comedic or dramatic effect: "scripted" has even become an adjective indicating an activity that seems spontaneous on one level but was planned in detail. The scripts may leave room for actors to improvise lines, or even record lines that were improvised on stage, but even ad-libbed lines do not reliably have the same features as truly spontaneous conversation.

- 24 Theatrical language also has a tendency to develop its own conventions and formulas, different from those of informal conversation. At regular intervals in the history of theater, playwrights and directors are acclaimed for having put "real language on stage," with the implication that the rest of theatrical language was unnatural.
- 25 There is a wide range of variation within theater, including genre variation (Wicks 1950 references comedy, drama, *folie*, pantomime and 28 other genres). The characters represented on stage are drawn from multiple social classes, genders and regions, and they are given voices that at least aim to represent the language of those classes, genders and regions.
- 26 Unfortunately, the characters in a corpus are not necessarily representative of the population that the corpus claims to present. There are several occasions where the corpus may be biased. The playwright chooses to focus on particular characters and to represent them in particular ways. The theater management chooses plays to produce based in part on what they expect will sell tickets. The theater may be censored by the government. Publishers choose plays to publish based in part on their success at the box office. Libraries and archives choose plays to store, to reproduce and to digitize. Corpus designers choose plays to include in the corpus.
- 27 It is impossible for us to account for all these potential sources of bias in a sampling frame. It is possible for us to control for some of them, and to be explicit with other corpus users about the remaining sources of bias. In the remainder of this paper I will discuss one source of bias in the FRANTEXT corpus, and the new Digital Parisian Stage corpus I am compiling that avoids that source of bias.

## 6. Bias in the FRANTEXT sampling frame

- 28 FRANTEXT is the primary corpus of nineteenth-century French, containing 76 plays from that century. It was originally compiled in the 1960s, as the basis for the *Trésor de la langue française* dictionary. The original intention was to create a representative sample, but the input and processing of texts took longer than anticipated, and to save time the selection process was changed (Imbs 1971: XXIII) to a "principle of authority." The corpus compilers consulted several well-regarded literary histories of French in the nineteenth and twentieth centuries, and made lists of every work mentioned in those histories. Nineteenth-century works that were mentioned five times or more and twentieth-century works that were mentioned four times or more were automatically included in the corpus. Works that were mentioned two or three times (or four times in histories of nineteenth century literature) were discussed in committee for possible inclusion, based on criteria such as their popularity, the richness of their vocabulary and the *sûreté* (roughly, confidence) of their language.
- 29 As I discussed above, the available archive of plays reflects the biases of the playwrights, theater managers, government censors, publishers, audiences, librarians, archivists and corpus designers. The principle of authority used in selecting texts for FRANTEXT

introduced another source of bias: the authors and publishers of the literary histories consulted. Not only were these authors biased by virtue of their social class and education, but the goals of these literary histories were to inform critics and writers about interesting developments in the literary canon, not to represent the full range of works produced during that period.

- 30 As we would expect from this bias, FRANTEXT is by no means a uniform sample. By text count, some authors are widely overrepresented: the list of 76 plays in the nineteenth-century segment of the corpus includes twelve plays by Théodore Leclerc, thirteen by Alfred de Musset and seven by Paul Claudel. The corpus for the period 1800-1815 contains four plays, which are listed in Appendix 1.
- 31 One of these four plays is a closet drama, a translation of Schiller's *Wallenstein* by Benjamin Constant (1809), which may never have been performed and does not appear to have been widely read outside of upper class intellectuals. The literary histories used for the Principle of Authority devote several pages to the discussion of the philosopher Germaine de Staël and her social circle. Constant was de Staël's romantic and intellectual partner for many years, and in those histories his work was discussed in that context.
- 32 This evidence of bias suggests that the texts in the FRANTEXT corpus will therefore not be representative of the input perceived by later playwrights. Is it possible to do better?

## 7. The Digital Parisian Stage (DPS) corpus

- 33 In order to improve on FRANTEXT, we need a more representative sample. For that, we will need a good sampling frame. We need a category of texts that relates to our research question, and an exhaustive list of texts in that category that we can use as the basis for our sample.
- 34 Fortunately, we have just such a list. From 1950 through 1979 Charles Beaumont Wicks compiled *The Parisian Stage*, a five-part catalog of every play that premiered in the French capital in the nineteenth century, from listings and reviews in contemporary newspapers, and secondary sources (Wicks 1950: vii). For the entire century, the total count of premieres is 31,879, although that includes several plays that Wicks listed in one part and then marked in subsequent parts as errors to be deleted.
- 35 I am currently in the process of compiling a new corpus, the Digital Parisian Stage, based on a representative sample of *The Parisian Stage*. Part 1 lists 3,017 plays from the years 1800 through 1815, and I have released a random 1% sample. Of these 31 listings, three are likely errors and six were never published. One of the six unpublished plays, the comic opera *Avis aux jaloux*, is available in manuscript form, from the archives of the Opéra Comique on [dezedede.com](http://dezedede.com).
- 36 One of the 22 published plays, *La Chaumière au pied des Alpes*, is available on microfilm as part of ProQuest's OmniSys World Literature Collection. The others are all available through Google Books, some with additional copies available from the Gallica website of the Bibliothèque Nationale de France, the Internet Archive and the University of Warwick's Marandet collection.
- 37 I obtained copies of all 23 available plays, processed them with ABBYY FineReader optical character recognition, and edited the resulting files to remove errors and provide consistent formatting and metadata.

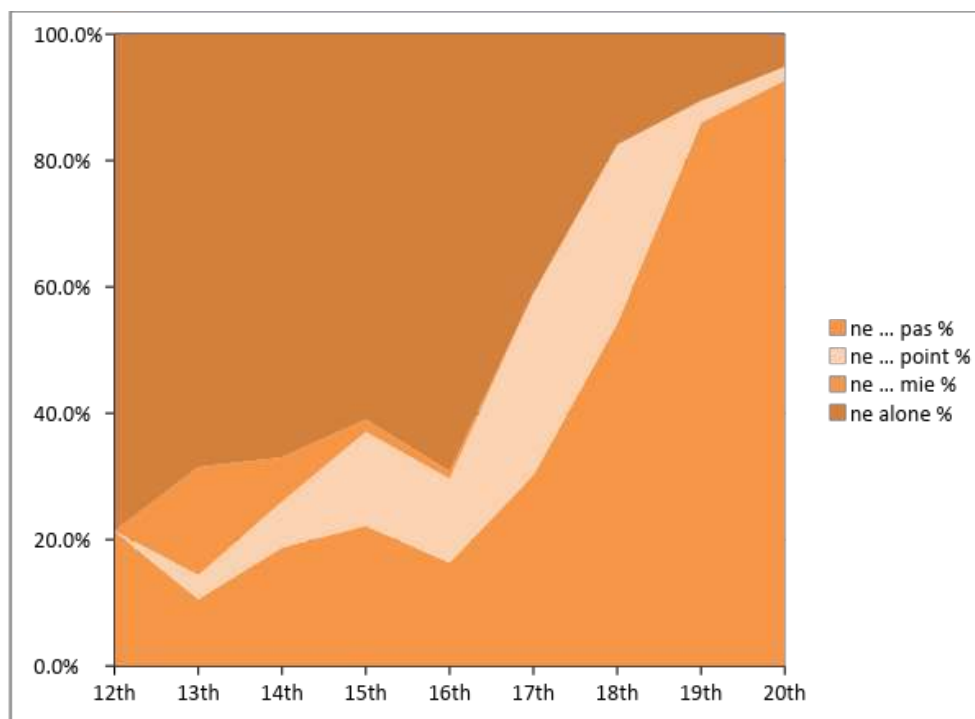
- 38 I have made the one percent sample list and the full texts of all 23 plays available on the software collaboration repository GitHub. In this choice I was inspired by authorship models that have been used in recent digital humanities literature, e.g. Tenen and Wythoff (2018). Sharing this corpus free of charge allows scholars to use it for any other purpose, including checking my work and running their preferred corpus analysis tools on the data. Publishing it on a collaborative platform like GitHub allows others to correct errors in the underlying texts, and to expand the corpus to include other texts.

## 8. Negation in FRANTEXT vs. the Digital Parisian Stage

- 39 As I discussed above, the research question in my Spread of Change study (Grieve-Smith 2009) was whether we can predict changes in the type frequencies of constructions in competition from one time period to the next based on their type frequencies in the earlier period. To test this question, I examined the evolution of negation in French. In particular, I focused on the shift in declarative sentence negation from *ne* alone to *ne ... pas* and *ne ... point*. To study this change I used a subset of the main reference corpus for French, FRANTEXT, and supplemented that with additional texts to cover periods before the beginning of the corpus in the seventeenth century. The general outline of the change is summarized in Chart 1:

Chart 1. The shift in declarative sentence negation from *ne* alone to *ne... pas* and *ne... point*

From Grieve-Smith (2009).



- 40 Here are examples of the three constructions used in plays from either FRANTEXT or the Digital Parisian Stage corpus:



(1)	LA DUCHESSE: et si vous <b>n'</b> en aviez de véritables [obstacles] à surmonter, où seroit la gloire de l'entreprise! (Pinto, 1800)
(2)	WALLSTEIN: Des fautes qu'il commit <b>n'</b> accusons que le sort. Il <b>n'est point</b> de courroux que <b>n'</b> apaise la mort. (Wallstein, 1809)
(3)	ANSELME: C'est être bien hardi, après toutes les menaces que vous avez osé me faire si je <b>ne</b> vous donnais <b>pas</b> ma fille... (Le Grenadier de Louis XV, 1815)
(4)	GUIGNOLET: jarni! je n'ons qu'un chagrin, c'est de <b>n'être pas</b> riche par nous-même; (Nitouche et Guignolet, 1802)

- 41 In the Spread of Change study I found evidence that suggested we could predict these changes in negation, using either a logistic model as proposed by Kroch (1988), or the more detailed competition models developed by Lotka (1925) and Volterra (1926), but that evidence was contingent on the assumption that the language measured for the first period (from the FRANTEXT corpus) was representative of the input perceived by the language users who produced the data for the second period (also in the FRANTEXT corpus).
- 42 The Digital Parisian Stage corpus allows us to test this assumption. Today the one percent sample is complete for the years 1800 through 1815 (Wicks, 1950). I annotated all 23 plays in this sample for declarative sentence negation, using an updated version of the custom PHP/MySQL application that I used in the Spread of Change study. I discarded one play, *La Mort du capitaine Cook*, because it was a pantomime with very little dialogue. I had already annotated one of the four plays in FRANTEXT from this period (*Cœlina, ou l'enfant du mystère*, for the Spread of Change study). I annotated the other three plays in the same way. The results for each play are available in Appendix 2; summary statistics are presented in Table 1 and Charts 1 and 2.

Table 1. Total token counts and average percentages of total declarative sentence negations expressed by the three negation constructions, 1800-1815.

	<i>ne</i> alone	<i>ne ... pas</i>	<i>ne ... point</i>	Total
FRANTEXT tokens	152	268	111	531
Digital Parisian Stage (DPS) tokens	234	1182	183	1599
FRANTEXT percent of total tokens	28.6%	50.5%	20.9%	
DPS percent of total tokens	14.6%	73.9%	11.4%	
Standard deviation of DPS percentage	0.126	0.212	0.160	
Student's <i>t</i> (of DPS relative to the FRANTEXT mean)	5.19	-5.18	2.77	

P	0.00002	0.00002	0.00578	
Cohen's <i>d</i>	1.11	-1.10	0.589	

Chart 1. Relative token frequencies of declarative sentence negation in the FRANTEXT corpus.

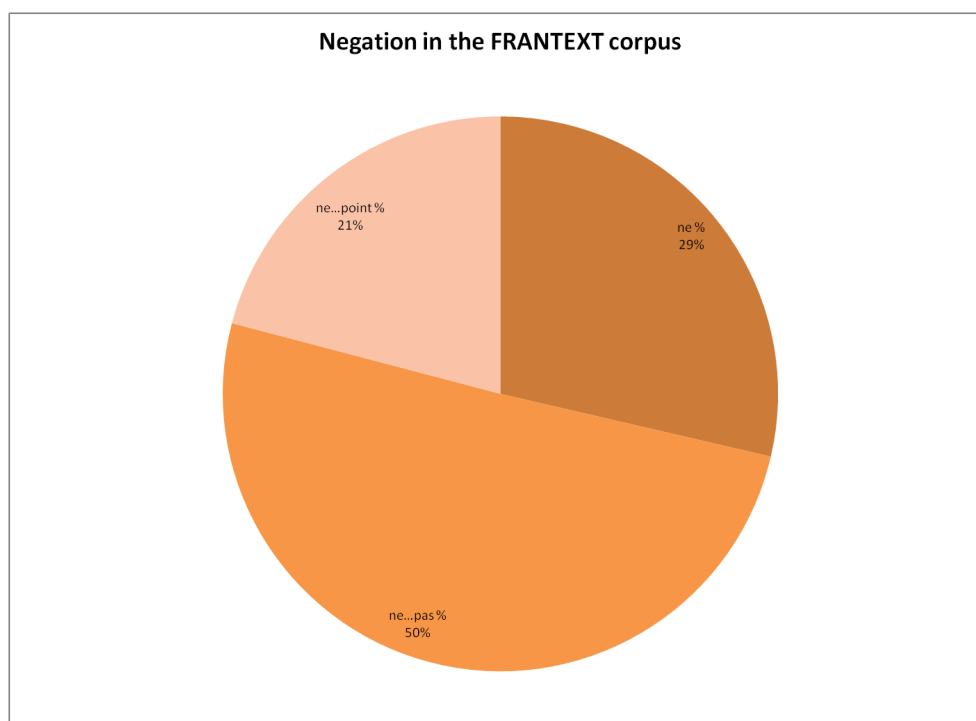
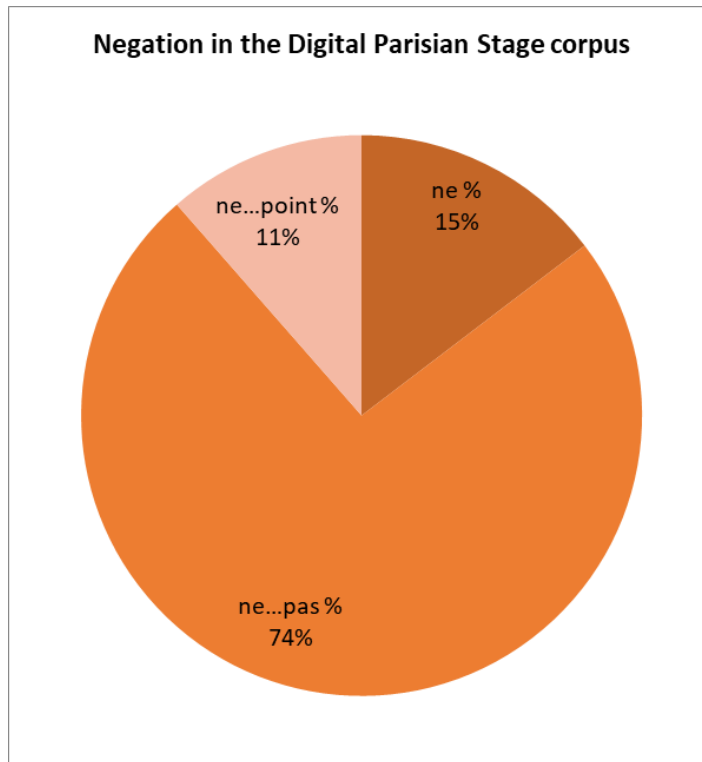


Chart 2. Relative token frequencies of declarative sentence negation in the Digital Parisian Stage corpus.



- 43 These results are striking. If we average the percentages for the four FRANTEXT plays, only slightly more than half the declarative sentence negations are made with *ne ... pas*. This is the type of statistic from FRANTEXT that is typically represented as what "French" was like at the time. But in the average Digital Parisian Stage play, almost three quarters of declarative sentence negations are made with *ne ... pas*. This yields a large effect (Cohen's  $d=1.1$ ).
- 44 There are two potential causes of the difference between these two corpora. The first is that the random sampling method happened to choose an unrepresentative subset of the texts. Student's (1908)  $t$ -test and Pearson's (1900) chi-squared ( $X^2$ ) test are designed to estimate just this type of error in a random sample. I have decided that I am comfortable if the chance of sampling error is less than one in twenty ( $\alpha=0.05$ ).
- 45 The  $t$ -test for the token frequencies of all three sentence negators yielded  $p$  values well below 0.05, meaning that we are justified in generalizing our results to the entire catalog of plays in Wicks (1950). The differences for the three negators are also statistically significant when tested together with the chi-square test ( $p < 0.05$ ;  $X^2=352$ ).
- 46 The second potential cause is that the absence of eight plays from the Digital Parisian Stage that were not available in full text may have biased the sample. As I mentioned in Section 7, three of those are likely errors, and the other five were never published. In the next section I will identify social factors that contribute to the difference in negation usage between the two corpora, and show that, based on these social factors, on average we would expect the five missing plays to be more innovative than the plays that were available. In other words, if at some point we find and analyze those plays, they are likely

to increase the difference in usage of the three negators between FRANTEXT and the Digital Parisian Stage, rather than decrease it.

## 9. Bias in FRANTEXT through the choice of theaters and genres

- 47 In early nineteenth century Paris, the vast majority of plays were identified in listings and in the title pages of printed scripts as belonging to a relatively small number of genres. The number of theaters had exploded when the royal monopoly was lifted in 1791, but many of those theaters had not lasted. Theaters were associated with particular genres, often explicitly. In 1807, Napoleon promulgated a series of decrees that restricted the number of theaters to four *grands théâtres* and four *théâtres secondaires*, and tied each theater to a repertoire of genres, by law. There were other venues, like the Cirque Olympique, that were licensed to provide different forms of entertainment, such as equestrian demonstrations, but many of them incorporated theatrical plots in an effort to attract audiences while skirting the restrictions on theatrical performances.
- 48 Different genres and theaters appealed to different social classes, and had a tendency to represent characters who resembled their audiences in many ways, including speech. Playwrights, theater directors, actors, censors, reviewers and spectators all worked together to establish patterns of style by genre and by theater.
- 49 Here are the four plays in the FRANTEXT corpus with their genres and theaters, and the frequency of sentence negators (also in Appendix 1):

Table 2. Genre, theater and the relative frequencies of declarative sentence negations in the four plays in the FRANTEXT corpus (1800-1815).

Title	Genre	Theater	<i>ne</i> alone %	<i>ne ... pas</i> %	<i>ne ... point</i> %
Coëlina, ou L'Enfant du Mystère	drame	Théâtre de l'Ambigu-Comique	18.49%	60.27%	21.23%
La Mort de Henri IV	tragédie	Théâtre Français	39.36%	48.94%	11.70%
Pinto	comédie héroïque	Théâtre Français	27.57%	52.97%	19.46%
Wallstein	tragédie	Closet	34.91%	33.96%	31.13%
Average			28.63%	50.47%	20.90%

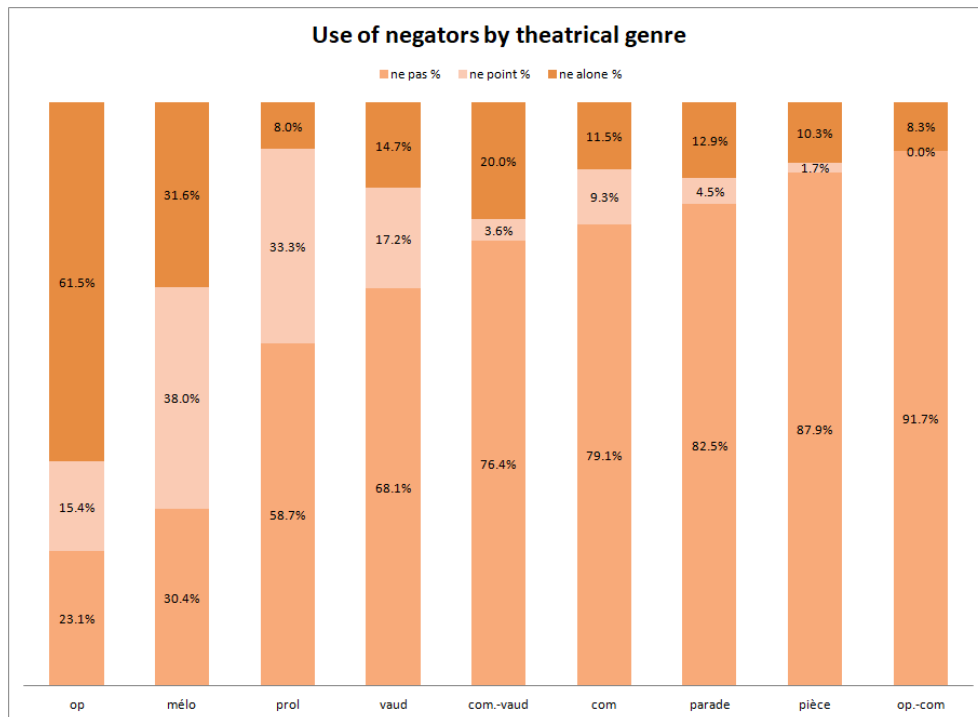
- 50 The following table shows how negation features are distributed across genres in the available plays of the Digital Parisian Stage corpus:

Table 3. Average relative frequency of declarative sentence negations by genre in the Digital Parisian Stage corpus.

Genre	Number of plays	<i>ne</i> alone%	<i>ne ... pas</i> %	<i>ne ... point</i> %
com	8	11.87%	79.08%	9.05%
com.-vaud	2	16.18%	78.63%	5.19%
mélo	1	31.65%	30.38%	37.97%
op	1	61.54%	23.08%	15.38%
op.-com	2	8.30%	91.70%	0.00%
parade	2	12.91%	82.55%	4.55%
pièce	1	10.34%	87.93%	1.72%
prol	2	8.00%	58.67%	33.33%
vaud	3	14.72%	68.10%	17.18%
$F(8,22)$		7.24	3.76	1.25
$p$		0.000978	0.0167	0.348
$\eta^2$		0.816	0.653	
$\omega^2$		0.694	0.468	

- 51 For *ne* alone and *ne ... pas*, one-way analysis of variance (ANOVA, Fisher 1921) allows us to rule out the possibility that the differences between genres are due to random bias in the sample. The following stacked bar chart shows the genres ordered by relative frequency of *ne ... pas*:

Chart 3. Average relative frequency of declarative sentence negations by genre in the Digital Parisian Stage corpus, sorted by relative frequency of *ne ... pas*.



- 52 Comparing the four FRANTEXT plays with the genre averages in the Digital Parisian Stage for negation, we see that *Cœlina* and *Pinto* are close to *Les Stréclitz*, the single melodrama and the second most conservative play in the Digital Parisian Stage sample. The two tragedies are more conservative than that, closer to the single opera in the sample, *Nephtali*. The other 21 plays in the sample all use more *ne ... pas* and less *ne alone* than any of the FRANTEXT plays.
- 53 Of the five missing plays in the Digital Parisian Stage, the only one that shares a genre with a FRANTEXT play is the melodrama *L'Abbaye de Grasville*. The other four include two comedies, a melodrama-comedy and a vaudeville. If these plays fit with the available comedies and vaudevilles in the corpus, we would expect them all to be more innovative in negation frequencies than the FRANTEXT plays, increasing the difference between the averages of the two corpora.
- 54 The following table shows how these negations were distributed across theaters in the Digital Parisian Stage corpus:

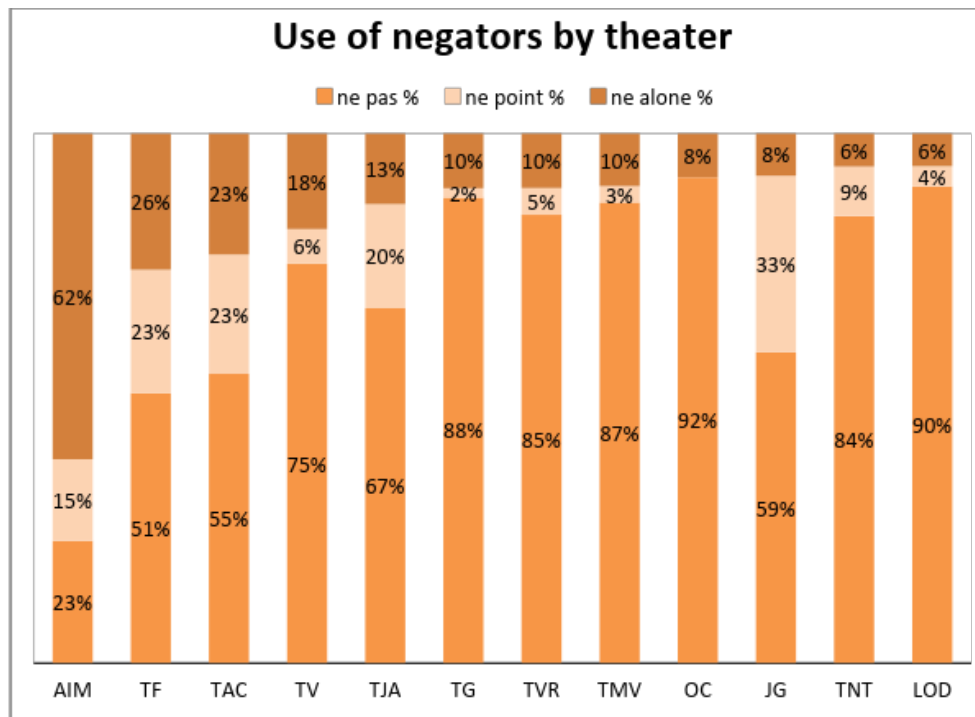
Table 4. Average relative frequency of declarative sentence negations by theater in the Digital Parisian Stage corpus.

Theater	Rank	Number of plays	<i>ne alone</i> %	<i>ne ... pas</i> %	<i>ne ... point</i> %
AIM	grand	1	61.54%	23.08%	15.38%
JG	tolerated	2	8.00%	58.67%	33.33%
LOD	grand	1	6.19%	90.00%	3.81%

OC	grand	2	8.30%	91.70%	0.00%
TAC	secondary	2	22.89%	54.58%	22.52%
TF	grand	1	25.73%	50.97%	23.30%
TG	secondary	1	10.34%	87.93%	1.72%
TJA	closed	3	13.33%	66.99%	19.68%
TNT	closed	1	6.25%	84.38%	9.38%
TV	secondary	3	18.06%	75.45%	6.49%
TVR	secondary	5	10.16%	85.57%	4.27%
$F(10,22)$			7.38	2.21	0.819
$p$			0.00136	0.0632	0.620
$\eta^2$			0.870		
$\omega^2$			0.744		

- 55 The relative frequency of *ne* alone was the only one of the three negators where the variance was great enough to rule out the possibility of sampling error, as determined by one-way ANOVA. The following chart illustrates these values in descending order by relative frequency of *ne* alone:

Chart 4. Average relative frequency of declarative sentence negations by genre in the Digital Parisian Stage corpus, inversely sorted by relative frequency of *ne* alone.



- 56 As with genre, the FRANTEXT plays are drawn from the repertoires of the more conservative theaters. *La Mort de Henri IV* and *Pinto* were performed at the Comédie-Française, the second most conservative theater for the use of *ne* alone. It is not surprising that *Wallstein*, as an elite closet drama, is even more conservative. *Cœlina* premièred at the Théâtre de l'Ambigu-Comique, the third-most conservative theater.
- 57 If we assume that each of the five missing plays resembled the other plays performed at the same theater, we can guess how they would have used language. *La Méprise* was performed at the Comédie-Française, and would likely be as conservative as *Pinto*. *La Dame invisible* was performed at the Théâtre du Vaudeville, and we would expect it to have moderate frequencies of *ne* alone. *L'Épreuve excusable* was performed at the Théâtre de la Gaîté, and would be as innovative as *Le Grenadier de Louis XV*. *L'Abbaye de Grasville* premièred at the Théâtre du Marais and *le Baron de Felsheim* at the Théâtre de la Porte-Saint-Martin. These theaters closed midway through the 1800-1815 sample period, and thus there are no plays from those theaters in the sample with available full text. Based on the first three of these plays, as with genres, we would expect the average negation frequencies to be more innovative if these five plays are found and added to the corpus.

## 10. Bias in FRANTEXT by age, gender and social class

- 58 Linguists have abundantly documented the effects of age (Labov 1963), gender (Trudgill 1972) and social class (Labov 1966). It would thus be normal to find effects of these variables in the relative frequency of sentence negation. To test this I focused on the lines of the one character in each play who produced the largest number of sentences, measured by periods, semicolons, question marks and exclamation points (with multiple



consecutive tokens of these counting as a single sentence delimiter). Unfortunately, none of these effects were strong enough to rule out sampling error.

- 59 It would be ideal to separate the speech of every character in each corpus, and look for effects of social class, age and gender, but this would introduce two potential confounding factors. First, it is well documented (Giles 1979) that people have a tendency to accommodate their speech to that of their interlocutors. Secondly, the characters in each play are the product of a single playwright or team of playwrights, sometimes in conjunction with an acting troupe. While most playwrights and players are clearly able to differentiate the speech of their characters within a given play, they may also impose stylistic uniformities that are not typical of spontaneous speech. It would be possible to test and compensate for these confounding factors in a follow-up analysis. For this study I chose to measure only the character with the highest sentence count.
- 60 Of these characters, only one was below adult age: Félix, the leader of the titular *Petits Braconniers*, who I classified as a Teen. I classified some as Elderly because they were described as *vieille* in the dramatis personae (Marcelline), had adult children (Maria) or took a parental role towards young adults (Hortence). Several characters were described as *jeune* (Sainville, Guignolet), or were newlyweds (Jocrisse-Maître), employed bourgeois living with their parents (Charles, Piron) or subjects of marriage plots where their age was unremarked (Jacques, Suzette), and therefore were classified as Young Adults. The rest were classified as Adults.

Table 5. Effects of character age on the relative frequency of constructions for declarative sentence negation in the Digital Parisian Stage corpus.

Age	Number of characters	<i>ne</i> alone%	<i>ne ... pas</i> %	<i>ne ... point</i> %
Adult	9	23.73%	66.27%	10.00%
Elderly	3	12.45%	77.84%	9.71%
Teen	1	0.00%	100.00%	0.00%
Young adult	9	13.40%	80.95%	5.65%
<i>F</i>		1.56	1.26	0.429
<i>p</i>		0.233	0.316	0.734

- 61 There were no explicitly gender variant or genderfluid characters, and the gender roles of the characters were often central to the plots. In *Papirius*, Calphurnie is the leader of an uprising of patrician women, and in *Les Mœurs du jour*, Formont's status as the older brother of Madame Dirval enables him to protect her from predatory men while her husband is at war. As is normal for French, every one of these characters was referred to with explicitly gendered pronouns and adjectives, which made them easy to classify.

Table 6. Effects of character gender on the relative frequency of constructions for declarative sentence negation in the Digital Parisian Stage corpus.

Gender	Number of characters	ne alone%	ne ... pas%	ne ... point%
Men	19	14.61%	75.06%	10.33%
Women	4	14.64%	78.15%	7.20%
p		0.993	0.676	0.600

- 62 To examine social class I took the occupational, family and class information provided in the *dramatis personae* of each play. Their individual negation usage is reported in Appendix 2. Some (Maria, Calphurnie) were described with explicit ranks as belonging to the nobility. Others were identified as domestic servants or slaves (Jacques, Esope). With the servants I also included Jocrisse-Maître, who had lived as a servant for all his life until just before the start of the play. There were characters who were explicitly described as peasants (Guignolet, Zimeline) or farm boys (Maurice). I divided the remaining characters into rural landowners (*propriétaires*, like Formont and La Grenade), and urban *bourgeois* (either Parisians in Paris like Charles Favart, or Parisians visiting the country like Sainville). Among the rural landowners I included Monsieur Tatillon because he successfully passes for a prosperous villager, despite not owning any property.

Table 7. Effects of character social class on the relative frequency of constructions for declarative sentence negation in the Digital Parisian Stage corpus.

Class	Number of characters	ne alone%	ne ... pas%	ne ... point%
Bourgeois	6	12.72%	80.24%	7.04%
Domestique	3	11.71%	78.47%	9.82%
Noble	6	21.15%	57.33%	21.52%
Paysan	3	14.70%	73.46%	11.84%
Propriétaire	4	12.74%	80.05%	7.21%
F		0.434	1.19	0.736
p		0.783	0.350	0.580

- 63 There appear to be patterns in the data that are not strong enough to rule out the possibility of sampling error. A follow-up study with a larger sample could potentially confirm some of these patterns. This might even be possible with a one percent sample of the full nineteenth century.

## 11. Implications for historical studies

- 64 One of the major assumptions underlying the Spread of Change study was that the texts in FRANTEXT for one period are representative of the language input for authors who wrote the texts in FRANTEXT in a subsequent period. The current finding, that there is a large difference between the use of negation in FRANTEXT and in the average Parisian play, invalidates that assumption. And yet the Spread of Change study found that the Lotka-Volterra model fit the FRANTEXT data. What are the implications of this for the Spread of Change study, and for future studies of historical syntax?
- 65 Using either corpus requires a layer of subjective interpretation. One possible assumption is that even if FRANTEXT is not representative of all Parisian plays, it is representative of the elite literary canon. The results of the Spread of Change study would then suggest that for the purposes of the theory of analogical extension, elite literary plays in one period serve as the input for elite literary playwrights of the next period. This in turn would mean that in French theater, elite literary genres are self-contained and norms are rigidly enforced.
- 66 Usage-based theories hypothesize that the dynamic of competition based on type frequency exists not just in elite literary genres, but everywhere in language. Have we now found evidence for it in theater, outside of elite literary subgenres? Not in this study, unfortunately. The Spread of Change study measured the effect at the scale of centuries, and found that the language of one century could predict the language of the following century. We may well see evidence of this effect at a level below that of centuries, but we are unlikely to find a measurable effect across a fifteen-year period.
- 67 Work is continuing on part II of the Digital Parisian Stage (1816-1830), and the rest of the nineteenth century will follow after that. It should be possible to test the Lotka-Volterra model on the full century, and possibly on half a century. This is an open source project that will be valuable to anyone studying changes in nineteenth-century Parisian French. It may take years with one scholar working to OCR, clean and format the texts, but if other scholars contribute to the process, the data will be available sooner.

## 12. Conclusion

- 68 Usage-based linguistic theories suggest that the language that people perceive affects the language that they produce. These theories are based on historical corpora, which we know to be biased in favor of elite literary language. I compared the language in the theatrical component of the FRANTEXT corpus with that in the Digital Parisian Stage corpus, a new corpus based on a representative sample of theater performed in the first sixteen years of the nineteenth century.
- 69 Comparing negation in the two corpora shows that the bias in the collection of FRANTEXT affects the structure of the language in the corpus. Declarative sentence negation in the Digital Parisian Stage plays was expressed with *ne ... pas* in 73.9% of tokens, on average, but the average FRANTEXT play used *ne ... pas* only 50.9% of the time.
- 70 It is not possible to sample the totality of the language that nineteenth-century speakers of French were exposed to over their lives. The language of theater, and possibly even the

language of elite theater, may be workable substitutes, but the selection methods used for corpora like FRANTEXT do not provide even a representative sample of elite theater.

- 71 The Digital Parisian Stage corpus is based on the work of Wicks (1950 et seq.), covering the entire nineteenth century. It currently covers Part I (Wicks 1950), and work is continuing on Part II (Wicks 1953). Once it is complete, it will allow linguists to study any aspect of the language of French theater in the nineteenth century, including changes that occurred during that century. The corpus is freely distributed on GitHub, and all scholars interested in the language of nineteenth-century Paris are encouraged to contribute to it.

---

## BIBLIOGRAPHY

- Bernoulli, Jacob. 1713. *Ars Conjectandi*. Basel : Thurnisiorum.
- Biber, Douglas. 1993. "Representativeness in corpus design." *Literary and Linguistic Computing* 8. 4.
- Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10. 425-55.
- Fisher, Ronald A. 1921. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, 1. 3-32
- Francis, W. Nelson, & Henry Kučera. 1964. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Providence : Brown.
- CNRTL. 2018. CNRTL : Centre National de Ressources Textuelles et Lexicales - Frantext, <http://www.cnrtl.fr/corpus/frantext/>
- Giles, Howard, and Philip Smith. 1979. Accommodation Theory : Optimal Levels of Convergence. In Giles and St. Clair, eds. *Language and Social Psychology*. Baltimore : Basil Blackwell.
- Grieve-Smith, Angus B. 2009. *The Spread of Change in French Negation*. Albuquerque : University of New Mexico dissertation.
- Grieve-Smith, Angus B. 2016. The Digital Parisian Stage. <https://github.com/grvsmth/theatredeparis>
- Imbs, Paul. 1971. *Trésor de la langue française*. Paris : CNRS.
- Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1. 199-244.
- Laplace, Pierre-Simon de. 1814. *Essai philosophique sur les probabilités*. Paris : Courcier.
- Labov, William. 1963. The Social Motivation of a Sound Change. *WORD* 19. 273-309.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington : Center for Applied Linguistics.
- Lodge, Anthony. 1991. Molière's peasants and the norms of spoken French. *Neuphilologische Mitteilungen* 92. 485-499.

Lotka, Alfred. 1925. *Elements of Physical Biology*. Baltimore : Williams and Wilkins.

Pearson, Karl. 1900. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" *Philosophical Magazine. Series 5*. 50. 157-175.

Student. 1908. The Probable Error of a Mean. *Biometrika* 6. 1-25.

Tenen, Denis, & Grant Wythoff. 2018. Sustainable Authorship in Plain Text using Pandoc and Markdown. *The Programming Historian*. <https://programminghistorian.org/en/lessons/sustainable-authorship-in-plain-text-using-pandoc-and-markdown>

Trudgill, Peter. 1972. Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich. *Language in Society* 1. 179-195

Volterra, Vito. 1926. Fluctuations in the abundance of a species considered mathematically. *Nature* 118.558-60.

Wicks, Charles Beaumont. 1950, 1953, 1961, 1967, 1979. *The Parisian Stage*. Tuscaloosa : University of Alabama.

Zager, David 1981. *A Real-time Process Model of Morphological Change*. Buffalo : State University of New York dissertation.

## APPENDIXES

### Appendix 1. Theatrical texts in FRANTEXT, 1800-1815.

Table 1. List of theatrical texts in FRANTEXT, (1800-1815), with authors and première dates.

Wicks number	Title	Author	Première date
582	Coëlina, ou l'enfant du mystère	Guilbert de Pixérécourt, René Charles	16 fructidor an VIII
2021	La Mort de Henri Quatre, Roi de France	Legouvé, Gabriel	June 25, 1806
2330	Pinto, ou la Journée d'une conspiration	Népomucène Lemerrier, Louis Jean	1 germinal an VIII
	Wallstein	Constant de Rebecque, Benjamin	1809

Table 2. Genre, theater and the relative frequencies of declarative sentence negations in the four plays in the FRANTEXT corpus (1800-1815).

Title	Genre	Theater	ne alone	ne ... pas	ne ... point

Cœlina, ou L'Enfant du Mystère	drame	Théâtre de l'Ambigu-Comique	18.49 %	60.27 %	21.23 %
La Mort de Henri IV	tragédie	Théâtre Français	39.36 %	48.94 %	11.70 %
Pinto	comédie héroïque	Théâtre Français	27.57 %	52.97 %	19.46 %
Wallstein	tragédie	Closet	34.91 %	33.96 %	31.13 %
Average			28.63 %	50.47 %	20.90 %

## Appendix 2. Texts in Part I of the Digital Parisian Stage corpus (1800-1815)

Table 1. All texts in the one percent sample, with titles, authors, première dates and availability.

Random rank	Wicks No.	Title	Première date	Author	Status
1	1229	le Grenadier de Louis XV, ou le Lendemain de Fontenoy	December 22, 1814	J.-B. Dubois	Available
2	1563	Jocrisse maître et Jocrisse valet	October 29, 1810	Sewrin	Available
3	1291	les Héritiers Michau, ou le Moulin de Lieursain	April 30, 1814	Planard	Available
4	125	un An de Périclès	January 1, 1810	J. Aude	Available
5	1178	Gallet, ou le Chansonnier droguiste	November 22, 1806	Moreau & d'Allarde	Available
6	2050	Nephtali, ou les Ammonites	April 15, 1806	E. Aignan	Available
7	2025	la Mort du Capitaine Cook, ou les Insulaires d'O-why-e	October 13, 1814	Franconi	Available
8	2701	le Singulier mariage	24 nivôse an IX	B. Dupont de Lille	Available
9	1903	la Méprise	November 22, 1815	Mme de Bawr	Not found

10	942	Esope chez Xantus	4 vendémiaire an IX	Tarenne de Laval	Available
11	926	l'Epreuve excusable	20 thermidor an VIII	Leroi de Neufvillette	Not found
12	2761	les Strélitz	May 12, 1808	Duperche & von Bilderbeck	Available
13	23	les Acteurs à l'épreuve	July 6, 1808	Sewrin & Chazet	Available
14	534	la Chaumière au pied des Alpes	May 4, 1810	Maillard, Brazier & Hapdé	Available
15	1127	Fontenelle	15 brumaire an XI	P.A.S. Petit aîné & Servières	Available
16	1544	la Jeunesse de Favart	February 11, 1808	A.-P.-C. Favart & Gentil de Chavagnac	Available
17	2576	le Rival obligeant	May 7, 1803	Mme de Bawr	Available
18	609a	le Compère futaille	February 18, 1803		Not found
19	1935	les Mœurs du jour, ou l'Ecole des jeunes femmes	7 thermidor an VIII	Collin d'Harleville	Available
20	371	la Bonne maîtresse, ou la Lettre trouvée	18 messidor an XI	Mme de Montanclos	Available
21	2861	les Trois Damis	January 4, 1804		Not found
22	259	Avis aux jaloux, ou la Rencontre imprévue	September 26, 1809	Chazet & Ourry	Available
23	2289	les Petits braconniers, ou les Ecoliers en vacances	April 5, 1813	Merle, Brazier & Charles Rondeau	Available
24	675	la Dame invisible	18 germinal an VIII	Chateaufvieux, Armand Croizette & Fleureau de Ligny	Not found
25	279	le Baron de Felsheim	March 26, 1805	Beaunoir	Not found
26	2065	Nitouche et Guignolet	17 nivôse an X		Available

27	2838	les Tracasseries, ou M. et Mme Tatillon	June 25, 1804	Picard	Available
28	12	l'Absinthe	January 14, 1805	Ch. Henrion	Available
29	5	l'Abbaye de Grasville, ou le fantôme imaginaire	February 19, 1804	Boirie et Clément	Not found
30	2341	Plus de peur que de mal	August 28, 1803		Not found
31	2169	Papirius, ou les Femmes comme elles étaient	11 messidor an IX	Gersin & Vieillard	Available

Table 2. Genres and theaters of the available plays in the Digital Parisian Stage corpus.

Wicks No.	Title	Genre	Theater
1229	le Grenadier de Louis XV	comédie	Théâtre de la Gaîté
1563	Jocrisse maître et Jocrisse valet	comédie	Théâtre Montansier-Variétés
1291	les Héritiers Michau	opéra-comique	Opéra-Comique
125	un An de Périclès	prologue	Jeux Gymniques
1178	Gallet	comédie	Théâtre Montansier-Variétés
2050	Nephtali	opéra	Académie Impériale de Musique
2025	la Mort du Capitaine Cook	pantomime	Cirque Olympique
2701	le Singulier mariage	vaudeville	Théâtre des Jeunes Artistes
1903	la Méprise	comédie	Théâtre Français
942	Esope chez Xantus	comédie-vaudeville	Théâtre du Vaudeville
926	l'Epreuve excusable	comédie	Théâtre de la Gaîté
2761	les Strélitz	mélodrame	Théâtre de l'Ambigu-Comique
23	les Acteurs à l'épreuve	vaudeville	Théâtre Montansier-Variétés
534	la Chaumière au pied des Alpes	prologue	Jeux Gymniques



1127	Fontenelle	vaudeville	Théâtre des Jeunes Artistes
1544	la Jeunesse de Favart	comédie-vaudeville	Théâtre du Vaudeville
2576	le Rival obligeant	comédie	Théâtre de l'Ambigu-Comique
609a	le Compère futaille	vaudeville	Théâtre du Marais
1935	les Mœurs du jour	comédie	Théâtre Français
371	la Bonne maîtresse	comédie	Théâtre des Jeunes Artistes
2861	les Trois Damis		Théâtre de la Société Olympique
259	Avis aux jaloux	opéra-comique	Opéra-Comique
2289	les Petits braconniers	parade	Théâtre Montansier-Variétés
675	la Dame invisible	vaudeville	Théâtre Montansier-Variétés
279	le Baron de Felsheim	mélodrame-comique	Théâtre de la Porte-Saint-Martin
2065	Nitouche et Guignolet	comédie	Théâtre Montansier-Variétés
2838	les Tracasseries	comédie	Théâtre de l'Odéon
12	l'Absinthe	comédie	Théâtre des Nouveaux Troubadours
5	l'Abbaye de Grasville	mélodrame	Théâtre de la Gaîté
2341	Plus de peur que de mal		Théâtre du Marais
2169	Papirius	parade	Théâtre du Vaudeville

Table 3. Token frequency of declarative sentence negation constructions in each available play in the Digital Parisian Stage corpus

Wicks No.	Title	<i>ne</i> alone	<i>ne ... pas</i>	<i>ne ... point</i>	Total negations
1229	le Grenadier de Louis XV	6	51	1	58
1563	Jocrisse maître et Jocrisse valet	8	63	7	78
1291	les Héritiers Michau	2	21	0	23
125	un An de Périclès	0	1	2	3
1178	Gallet	7	89	5	101

2050	Nephtali	8	3	2	13
2701	le Singulier mariage	15	50	34	99
942	Esope chez Xantus	11	72	6	89
2761	les Strélitz	25	24	30	79
23	les Acteurs à l'épreuve	8	37	3	48
534	la Chaumière au pied des Alpes	4	21	0	25
1127	Fontenelle	9	56	8	73
1544	la Jeunesse de Favart	11	42	2	55
2576	le Rival obligeant	14	78	7	99
1935	les Mœurs du jour	53	105	48	206
371	la Bonne maîtresse	10	59	11	80
259	Avis aux jaloux	3	35	0	38
2289	les Petits braconniers	2	48	0	50
2065	Nitouche et Guignolet	11	73	1	85
2838	les Tracasseries	13	189	8	210
12	l'Absinthe	2	27	3	32
2169	Papirius	12	38	5	55

Table 4. The most frequent character, with social class, age and gender, in each available play in the Digital Parisian Stage corpus.

Wicks No.	Title	Character	Class	Age	Gender
1229	le Grenadier de Louis XV	Francœur	Propriétaire	Adult	Man
1563	Jocrisse maître et Jocrisse valet	Jocrisse-Maître	Domestique	Young adult	Man
1291	les Héritiers Michau	Suzette	Propriétaire	Young adult	Woman
125	un An de Périclès	Méonidas	Noble	Adult	Man

1178	Gallet	Piron	Bourgeois	Young adult	Man
2050	Nephtali	Nephtali	Noble	Adult	Man
2701	le Singulier mariage	Maurice	Paysan	Young adult	Man
942	Esope chez Xantus	Esope	Doméstique	Adult	Man
2761	les Stréclitz	Maria	Noble	Elderly	Woman
23	les Acteurs à l'épreuve	Dupré	Bourgeois	Adult	Man
534	la Chaumière au pied des Alpes	Zimeline	Paysan	Young adult	Woman
1127	Fontenelle	Fontenelle	Bourgeois	Adult	Man
1544	la Jeunesse de Favart	Charles	Bourgeois	Young adult	Man
2576	le Rival obligeant	Sainville	Bourgeois	Young adult	Man
1935	les Mœurs du jour	Formont	Propriétaire	Adult	Man
371	la Bonne maîtresse	Jacques	Doméstique	Young adult	Man
259	Avis aux jaloux	Lorenzo	Noble	Elderly	Man
2289	les Petits braconniers	Félix	Noble	Teen	Man
2065	Nitouche et Guignolet	Guignolet	Paysan	Young adult	Man
2838	les Tracasseries	M. Tatillon	Propriétaire	Adult	Man
12	l'Absinthe	Hortence	Bourgeois	Elderly	Woman
2169	Papirius	Calphurnie	Noble	Adult	Woman

Table 5. Token frequency of declarative sentence negation constructions for the most frequent character in each play of the Digital Parisian Stage corpus

Wicks No.	Title	Character	<i>ne</i> alone	<i>ne ... pas</i>	<i>ne ... point</i>	Total negations
1229	le Grenadier de Louis XV	Francœur	6	51	1	58

1563	Jocrisse maître et Jocrisse valet	Jocrisse-Maître	8	63	7	78
1291	les Héritiers Michau	Suzette	2	21	0	23
125	un An de Périclès	Méonidas	0	1	2	3
1178	Gallet	Piron	7	89	5	101
2050	Nephtali	Nephtali	8	3	2	13
2701	le Singulier mariage	Maurice	15	50	34	99
942	Esope chez Xantus	Esope	11	72	6	89
2761	les Strélitz	Maria	25	24	30	79
23	les Acteurs à l'épreuve	Dupré	8	37	3	48
534	la Chaumière au pied des Alpes	Zimeline	4	21	0	25
1127	Fontenelle	Fontenelle	9	56	8	73
1544	la Jeunesse de Favart	Charles	11	42	2	55
2576	le Rival obligeant	Sainville	14	78	7	99
1935	les Mœurs du jour	Formont	53	105	48	206
371	la Bonne maîtresse	Jacques	10	59	11	80
259	Avis aux jaloux	Lorenzo	3	35	0	38
2289	les Petits braconniers	Félix	2	48	0	50
2065	Nitouche et Guignolet	Guignolet	11	73	1	85
2838	les Tracasseries	M. Tatillon	13	189	8	210
12	l'Absinthe	Hortence	2	27	3	32
2169	Papirius	Calphurnie	12	38	5	55

## ABSTRACTS

Cognitive linguistics makes specific claims about language use, and corpora are our most powerful tool to test those claims. Representative sampling (Laplace 1814) is a technique that allows us to study smaller, more manageable corpora, and generalize our results to a broader sampling frame. For a sampled corpus to be relevant to our research questions, its sampling frame must have an understandable connection to the subject of our research question.

In my dissertation study (Grieve-Smith 2009) I tested the type frequency hypothesis of analogical

extension (Bybee 1995) using the FRANTEXT corpus (CNRTL 2018). In this study I test the theatrical texts in FRANTEXT from 1800-1815 against the new Digital Parisian Stage corpus, sampled from Wicks (1950 et seq.), a catalog of every play that premiered in Paris in the nineteenth century. Declarative sentence negations in the Digital Parisian Stage corpus occurred with *ne ... pas* in 73.9 % of tokens, while in FRANTEXT they only occurred with *ne ... pas* in 50.5 % of tokens. This shows that FRANTEXT is biased in favor of elite literary language. To properly test usage-based theories of language change we will need a representative corpus covering a century or more.

Pour tester les hypothèses avancées par la linguistique cognitive, il n'y a pas d'instrument plus efficace que le corpus. L'échantillonnage représentatif (Laplace 1814) est une technique qui permet d'examiner des corpus plus réduits, et ainsi plus abordables, et d'en généraliser les résultats à un cadre d'échantillonnage plus large. Or, un échantillon n'est pas pertinent à une hypothèse s'il n'est pas tiré d'un cadre d'échantillonnage qui soit lui-même pertinent à l'hypothèse.

Dans mon projet doctoral (Grieve-Smith 2009) j'ai employé le corpus FRANTEXT (CNRTL 2018) pour tester l'hypothèse selon laquelle l'extension analogique d'une construction dépend de sa fréquence de type (Bybee 1995). J'ai comparé les textes théâtraux dans FRANTEXT pour les années 1800-1815 avec un nouveau Corpus de la Scène Parisienne, un échantillon tiré du catalogue de Wicks (1950 et seq.). Dans ce nouveau corpus, les négations de phrase déclarative se forment avec *ne ... pas* dans 73,9 % des instances, tandis que dans FRANTEXT elles ne se forment avec *ne ... pas* que dans 50,5 % des occurrences, une différence qui montre un biais en faveur de la langue littéraire dans FRANTEXT. Pour une évaluation adéquate des théories basées sur l'usage concernant le changement linguistique, il faudra un corpus contenant des textes représentatifs de la langue sur un siècle au minimum.

## INDEX

**Keywords:** Language change, French language, corpus design, sampling, usage-based, type frequency, analogical extension

**Mots-clés:** Evolution linguistique, structuration de corpus, échantillonnage, théories basées sur l'usage, fréquence de types, extension analogique

## AUTHOR

ANGUS B. GRIEVE-SMITH

The New School