

PhD Dissertation

Epidemiology of Representations: An Empirical Approach

—original title may change—

Sébastien Lérique¹

Supervisor: Jean-Pierre Nadal²
Co-supervisor: Camille Roth³

¹Centre d'Analyse et de Mathématique Sociales (CAMS, UMR 8557, CNRS-EHESS, Paris). Email: sebastien.lerique@normalesup.org.

²CAMS, and Laboratoire de Physique Statistique (LPS, UMR 8550, CNRS-ENS-UPMC-Univ. Paris Diderot, Paris). Email: nadal@lps.ens.fr

³CAMS, Centre Marc Bloch (CMB, UMIFRE 14, CNRS-MAEE-HU, Berlin), and Sciences Po, médialab (Paris). Email: camille.roth@sciencespo.fr

Contents

1	Introduction	5
2	Brains Copy Paste	7
3	Gistr	9
3.1	Introduction	9
3.2	Related work	9
3.3	Methods	10
3.3.1	Experiment design principles and process	10
3.4	Results	16
3.5	Discussion	16
4	Discussion	19
5	Conclusion	21
	References	23

Chapter 1

Introduction

Chapter 2

Brains Copy Paste

Chapter 3

Gistr

3.1 Introduction

3.2 Related work

- Bartlett
- Nettle
- Bebbington
- Mesoudi & Whiten
- Kirby & Tamariz
- Acerbi
- Claidière
- Smartphone psychology
- BCP
- *Missing information*: most blog and media outlet authors do not quote their source when they publish a quote online (it's often not relevant to the article), meaning there are no source-destination links in the data collected; this information must instead be inferred anew to study the evolution of content. There is also no access to author information (gender or age, experience in writing, but also psychological factors like memory span), ruling out any study of individual author effects in transforming the content.
- *Missing context*: the lack of access to the context of production and reception of quotes makes it impossible to interpret what a quotation means to its author or its reader (???; ???; ???). Analysing any kind of semantic evolution is therefore out of reach for this kind of passively collected online data (???, to be published).

their questions, and their limitations

3.3 Methods

3.3.1 Experiment design principles and process

Advantages and challenges of transmission chain

An obvious way to address the questions raised in the previous chapter is to use a transmission chain experiment to study the evolution of online quotations in a controlled setting: each subject reads, retains, and rewrites sentences that are then passed on to the next subject in a chain of reformulations. If designed carefully, such a setup can reproduce an idealised version of the read-remember-rewrite process which, we hypothesised, participates in the evolution of quotations in blogspace and media outlets. It also provides the information that our previous data set lacked in order to analyse the complete transformations of quotations, as well as the long-term effect of those changes: the links between parent and child sentences are naturally encoded in the data, such that the transformations undergone by each sentence can be studied in full detail without restricting ourselves to simpler changes as was necessary for the inference procedure used with digital traces from blogspace. By creating an artificial setting, the experiment design also lets us control the reading and writing conditions as well as the context in which sentences appear, which further removes one of the inevitable uncertainties of the previous protocol (albeit at the cost of less ecological conditions).

As previously, our approach is exploratory, which renders the laboratory transmission chain paradigm difficult to apply directly: we aim to collect data that will allow us to study both the complete set of transformations undergone by short utterances such as online quotations, and the interactions and cumulative effect of such changes; yet we do not know in advance the types of changes that subjects will make, or the extent to which such changes vary according to the type of linguistic content. Transmission chain studies typically start with an *a priori* hypothesis focused on a well-identified type of content, which is then empirically tested by contrasting the evolutionary outcome of two classes of sentences. Our goal here, however, is to provide first steps in characterising the process by which such evolution of linguistic content arises, and observe how it accumulates in the long term. The setup must thus allow us to collect enough data to extract regularities in successive transformations operated by different subjects on different sentences, and provide a resolving power similar to that of substitutions in online quotations so that we can compare results with the previous chapter. Since our main target is the set of detailed transformations and their interactions, versus targeting the contrast of accumulated outcomes, it is also crucial to fine-tune the difficulty of the read-write task and the complexity of the source sentences, in order to trigger a set of transformations varied enough that it could approach some of the changes encountered in real life situations. Our progress therefore involves a non-trivial trial-and-error component: indeed, a task made too easy or too difficult, and more so a set of source sentences that are too complex or too straightforward, will lead to either mass deletions or perfect conservation (or the former followed by the latter), none of which can help characterise the more intricate changes that linguistic content undergoes in the ecological setting we aim to simulate.

Web and smartphone experiments

Complementary to laboratory studies and to approaches using online digital traces, a new empirical approach based on Web browsers and mobile computing is striking a different balance in the trade-offs of experimental work which seems very promising in addressing the problems outlined above.

Indeed, browsers (both on desktop and mobile) and smartphones have evolved into powerful, ubiquitous application environments for which one can develop any kind of experiment involving text, graphics, and human interactions. At the cost of increased engineering requirements and a different approach to subject recruitment, Web and smartphone experiments give the designer full control over what data is collected and the way interactions are framed (similar to laboratory experiments), and make it possible to quickly collect data sets at scales comparable to what filtered and cleaned digital traces provide.

This approach makes a number of unusual trade-offs, the benefits of which can be summarised as follows:

- *Control*: similar to laboratory experiments, and unlike digital trace analysis, it is possible to use complex designs where all the interactions of the subjects are framed and observed by the experimenter. This includes for instance the presentation of the experiment (e.g. as a game or a self-improvement aid, aside from being a scientific study) and, more importantly, the ways in which the system mediates the interactions between the subjects.
- *Scale*: if and when needed, the technical platform can scale the number of subjects to the tens of thousands at low marginal cost. Interactions between subjects can also scale to involve synchronous or asynchronous contact between hundreds of people, without having to manage per-subject scheduling.
- *Speed of data collection*: once the initial development is completed (see costs below), the data collection cycle is short. One day can be enough to collect a data set comparable in size and quality to the final substitutions set extracted and analysed in the previous chapter. This is especially relevant for exploratory work which is made much easier with shorter trial-and-error cycles.
- *Flexible recruitment*: while also a challenge (see costs below), subject recruitment is more flexible than in the laboratory: services like Prolific Academic¹ let the experimenter recruit at reasonable costs in pools of tens of thousands of subjects with fine-grained demographic filters. Wider audiences can be achieved by offering non-financial rewards, framing the experiment as a self-improvement application, or turning it into a game.

The corresponding costs are the following:

- *Technical challenge*: developing Web and smartphone experiments involves a substantial amount of engineering, and makes use of an array of technologies that most researchers, even technical, are not familiar with. While some all-in-one kits exist,² creating an experiment that meets one's research questions often requires learning average skills in most of the various technologies at play: a native or cross-platform smartphone development environment, Web application development, backend server programming, and some server administration skills. Most importantly, the paradigms and problems encountered are often new to researchers: most programming is asynchronous due to network communication or the user interface, and technicalities such as user management or email validation can grow into difficult engineering challenges.
- *Spam-control*: subjects are not constrained or encouraged by the face-to-face interaction of a laboratory experiment, neither are they (in most experiments) in the course of an interaction with friends that provides natural incentives for what they write, as can be the case with digital traces. Participants must have an incentive to perform the experiment's tasks well. If the spam introduced by one subject can be isolated in the design of the experiment, one possibility is to filter it once the data is collected and make payment depend on its prevalence. But if the

¹<https://www.prolific.ac/>

²See e.g. <http://funf.org/> and <http://www.epicollect.net/>.

spam introduced by one subject naturally propagates to data seen by other subjects in the experiment, as is the case for transmission chains, effective anti-spam pressures and motivations need to be factored into the design.

- *Recruitment cost:* while recruiting a few dozen or even a few hundred subjects is generally cheaper (not counting the cost of developing the experiment) than the equivalent for a laboratory experiment,³ and can be very easy for fast prototyping and pilots, recruitment cost rises linearly with the number of subjects and the time they spend on the experiment, unless a different strategy is used. Turning an experiment into a playful application or an application useful to the subjects (effectively making them users) involves yet another set of skills, can prove challenging, and must be factored into the development cost.

Web-based transmission chains

The balance achieved by the Web-based approach is particularly adapted to the experimental requirements we outlined above. Since no existing system would fit our needs, we chose to develop a reusable Web-based platform to run a series of studies that would allow us, once ready, to gather sufficient amounts of quality data in short cycles. We further decided to implement the simplest possible version of the transmission chain paradigm that is still viable, and leave the exploration of more complex setups for future research: the task we used asks subjects to read and memorise a short utterance, wait a few seconds, then rewrite what they have read as accurately as possible.

Subjects' productions are arranged in chains, such that what a given subject produces is used as the source utterance for the subject appearing next in a chain. In particular, the utterances to memorise are presented with no surrounding context, no distraction task is used between the reading and writing phases, and the material incentive for the task is purely monetary (although as we describe below we fine-tuned the interface to strongly encourage subjects to be conscientious). This simple setup lets us quickly gather data sets of several thousand utterance transformations, ensuring our results will be comparable to those from the set of 6177 substitutions we extracted in the previous chapter. Two parameters are left to vary: the reading time for the source utterances, computed as the number of words in an utterance multiplied by a reading factor that is to be adjusted, and the set of initial source utterances.

The experiment is available to subjects as a website, and passing it involves the following steps:

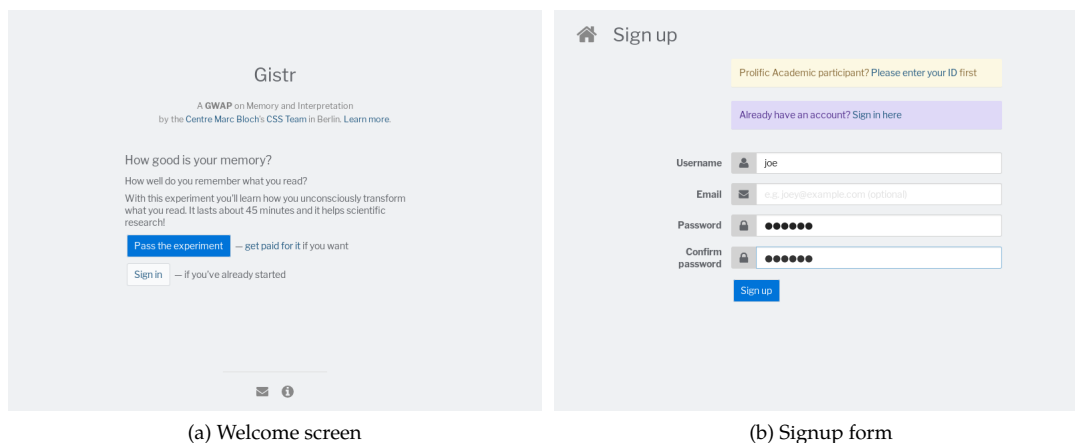
- Welcome and sign up (Figs. 3.1a and 3.1b),
- Answering a preliminary questionnaire (Fig. 3.2),⁴
- Subjects then start training for the main task, where they are asked to repeatedly memorise and rewrite short utterances as accurately as possible. As the instructions in Fig. 3.3 indicate, an utterance is presented to the subject and after a short pause they are asked to rewrite it as remembered. The process loops until the subject has completed all the utterances assigned to them (calibrated so that completing the experiment lasts at most one hour). The real trials start after 3 to 5 training trials, depending on the overall experiment length.

Each utterance from the initial set is used to create several parallel chains in order to allow for comparisons across chains with the same initial utterance. The final data thus consists in a set of reformulation trees, where each tree branch is a transmission chain started from the tree root, and

³Global competition on online platforms like Prolific Academic drive subject payments down.

⁴An early version of the experiment also included a word span test at this stage. However, similarly to the age of subjects that we collect in the questionnaire, this data turned out to not be relevant in the analyses. The magnitude of transformations depends far more on the conscientiousness of subjects, and this non-trivial test was later removed during one of the frontend rewrites.

continuing until it reaches a target depth defined for the experiment.⁵ The number of branches in a tree is also adjusted for each run of the experiment. Except for those who drop out before finishing the experiment, all subjects are exposed exactly once to each tree in random order, such that all the reformulations in a given tree are made by distinct subjects, and nearly all subjects (excluding dropouts) are present in each tree. Finally, note that when exposed to a tree, subjects are always randomly assigned to the tip of one of the branches that have not yet reached the target depth: subjects are thus randomly distributed across branches, but their depth-ordering loosely corresponds to time of arrival on the tree. In particular, if a subject starts the experiment after most other subjects have completed it, he or she will be mostly exposed to utterances deep in the branches. Due to the chained nature of the data, there is no economical way of countering this ordering bias.⁶ Fig. 3.4 shows a representation of the shape of the final trees.



(a) Welcome screen

(b) Signup form

Figure 3.1: Initial steps for a subject entering the experiment.

Technically, the platform is a complete Web application based on current technologies, with accompanying backend server to collect and distribute utterances.⁷ The experiment is available at <https://gistr.io> and subject recruitment was done using Prolific Academic, a service analogous to Amazon Mechanical Turk and geared towards academic research.⁸

Using the Prolific Academic service allowed us to select among a pool of over 26,000 subjects, for

⁵We therefore use the terms “chain” and “branch” interchangeably in what follows.

⁶The following three approaches could be combined to counter ordering bias. (1) Have each subject do a single trial, that is, use as many subjects as there are reformulations in the full experiment; this is extremely expensive as there is a fixed minimal price for each subject, in order to give time to explore the interface, answer the initial questionnaire, and train for the task. (2) Have each subject wait an adjustable amount of time between each trial, to open the possibility of ordering subjects differently from their time of arrival; this is also expensive, as it means paying subjects for waiting most of the time they spend on the experiment. (3) Optimise the order of tree presentations of each subject so as to spread subjects across depths; while this approach could achieve some level of spread when combined with (2), it is contingent on the starting times of subjects and their synchronisation, which we do not control (subjects find the experiment through Prolific Academic notifications and are free to start whenever they want).

⁷The frontend first used the Ember.js framework (Ember.js contributors 2017), and was later rewritten and extended using the Elm programming language (Czaplicki and Elm contributors 2017). The backend is a Python application written on top of the Django REST framework (Christie and Django REST framework contributors 2017). Most of the critical logic in the software is verified using automated tests, and the full source code is available under a Free Software licence at <https://github.com/interpretation-experiment/gistr-app> (frontend), and <https://github.com/interpretation-experiment/spreadr> (backend).

⁸The public url of the experiment was not advertised anywhere else, and checking the subjects’ Prolific Academic ID confirmed that only people from that platform participated in each experiment.

Profile Signed in as **joe** — [Sign out](#)

Dashboard
Settings
Emails

General questionnaire

We'd like to know a bit about you before you start the experiment. This will help us understand what influences your results as well as other participants' results.

Your answers will be kept strictly private and will only be used for the purposes of the experiment.

It takes about 2 minutes to fill the questionnaire. Thanks for participating, and welcome again to Gistr!

About you

Age

Gender
☐ Female ☐ Male ☐ Other

☐ Check this if you know what this experiment is about

Your schooling and what you do

We'd like to know how much you've studied, as well as what type of job you work in, or what your main daily activity is.

What is the highest level of education you attained?

Please select from the list

Please describe, in your own words, the highest level of education you attained. You can use several sentences if necessary.

What is your general type of profession or main daily activity?

Please select from the list

Please describe your profession or main daily activity. You can use several sentences if necessary.

Is there something wrong with this questionnaire, or a comment you would like to share? Please [tell us about it!](#)

Feedback

Figure 3.2: Initial questionnaire. Subjects can additionally submit feedback on the questionnaire or any other aspect of the experiment on most screens of the website.

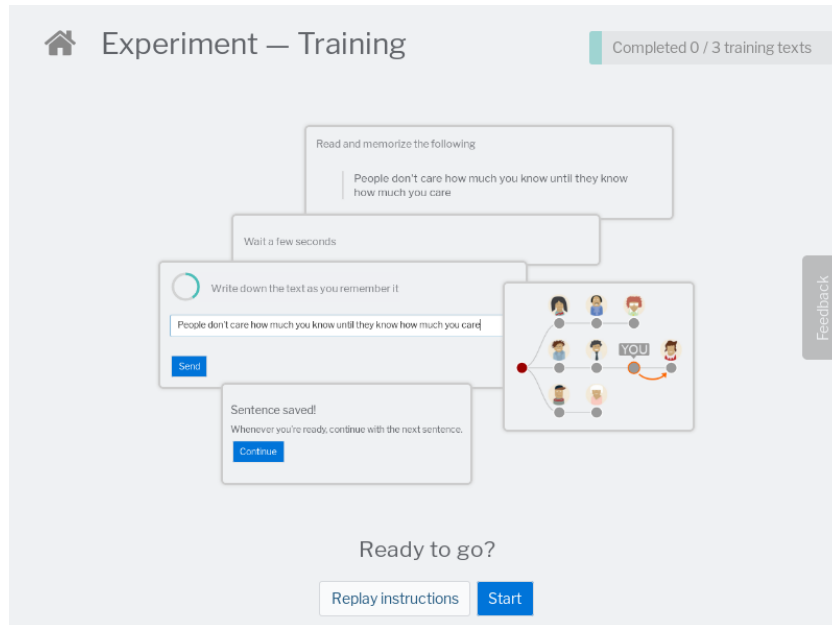


Figure 3.3: Instructions for the main task

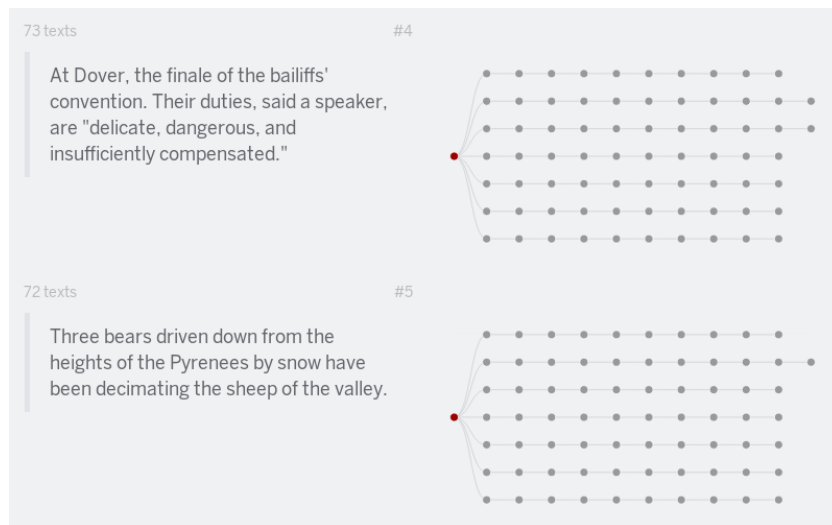


Figure 3.4: Two example reformulation trees generated by the setup: the text on the left is the initial utterance for all branches of a tree, represented by a red dot in the right-hand graph; each grey dot in the graph represents an utterance produced by a subject on the basis of the preceding dot. Most subjects created one reformulation in each tree; however, since subjects from Prolific Academic do not always complete the whole set of utterances assigned to them, we recruit additional subjects to fill the trees that were left incomplete. This leads the other, already complete trees to receive more reformulations than needed, making some of their branches run deeper than the target depth (as is the case for the trees shown here). All branches are cropped to the target depth for analysis.

which we used the following criteria:

- First language English speaker
- At least 18 years old
- Current country of residence and place of most time spent before turning 18 must both be the UK
- Normal or corrected to normal vision
- No diagnosed literary difficulties
- Completed secondary school
- Not having participated in any of the preceding experiments

Only the first two constraints were enforced for the first experiment, and the full set was used for all subsequent runs. The full filter provided over 2300 eligible subjects, from which the service automatically sampled the number of subjects we requested.

3.4 Results

3.5 Discussion

The need for embedding

Any quantitative study relies on abstracting out details of particular cases by reducing (most often averaging) values in each dimension to a few indicators. Being able to render a precise view of the studied phenomenon then depends on being able to determine which are the right dimensions to describe it, and having access to them (???).

Embedding experiments in the everyday life of subjects gives access to dimensions that can be otherwise unavailable: through the use of smartphones, an experiment designer can trigger interactions with subjects (for instance asking questions) at any moment of the day, or have measures running while subjects are offline (using the ever-increasing number of sensors present on the devices), both of which are impossible with digital traces. Above all, embedding an experiment means getting greater access to context, which opens the possibility of understanding phenomena the way they are meant in the lives of subjects, and not only in the way they are construed by the experiment designer.

Daydreaming is an example experiment developed as a smartphone application with Vincent Adam, Mikaël Bastian, Jérôme Sackur, and Gislain Delaire, that took advantage of this embedding. The experiment, focused on our awareness of daily mind-wandering, would probe subjects during a month at random moments of the day to ask them if they were mind-wandering (and, if so, what were the qualities of their thoughts).⁹ While our team spent over a year developing the application, it allowed asking questions related to ecological situations which cannot exist in laboratory or passive collection studies. [section 3.5](#) shows a sample question asked to the subjects, and [section 3.5](#) shows an example of the results produced at the end of the experiment (this particular screen shows the results for one subject; seeing their own results was part of the reward for subjects participating in the study).

Example questionnaire in the Daydreaming app

⁹See <http://daydreaming-the-app.net/> for more details.

Results on weekly mind-wandering rhythms from the Daydreaming app

- *Embedding*: as explained above, smartphone-based experiments allow for real-life embedding: the experiment designer can choose when and how interactions with the experiment and between subjects take place, and measure any number of variables the device gives them access to (geolocation, time, phone agitation through its accelerometers, general noise level, etc.), virtually at any moment.

Chapter 4

Discussion

Chapter 5

Conclusion

References

Christie, Tom, and Django REST framework contributors. 2017. 'Django REST Framework'. <http://www.django-rest-framework.org/>.

Czaplicki, Evan, and Elm contributors. 2017. 'Elm: A Delightful Language for Reliable Webapps'. <http://elm-lang.org/>.

Ember.js contributors. 2017. 'Ember.js: A Framework for Creating Ambitious Web Applications.' <https://emberjs.com/>.