# Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change 

Q P Feltgen, Benjamin Fagard, Jean-Pierre Nadal

## To cite this version:

Q P Feltgen, Benjamin Fagard, Jean-Pierre Nadal. Frequency patterns of semantic change: Corpusbased evidence of a near-critical dynamics in language change. 2017. hal-01652107v1

HAL Id: hal-01652107
https://shs.hal.science/hal-01652107v1
Preprint submitted on 6 Mar 2017 (v1), last revised 29 Nov 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

# Frequency patterns of semantic change: Corpus-based evidence of a near-critical dynamics in language change 

Q. Feltgen ${ }^{1}$, B. Fagard ${ }^{2}$ and J.P. Nadal ${ }^{1,3}$<br>${ }^{1}$ Laboratoire de Physique Statistique, École Normale Supérieure, PSL Research University; Université Paris Diderot, Sorbonne Paris-Cité; Sorbonne Universités, UPMC - Univ. Paris 06; CNRS; Paris, France.<br>${ }^{2}$ Laboratoire Langues, Textes, Traitements informatique, Cognition (LaTTiCe, UMR 8094 CNRS - ENS - Université Paris 3), École normale supérieure, Paris, France.<br>${ }^{3}$ École des Hautes Études en Sciences Sociales, PSL Research University, CNRS, Centre d'Analyse et de Mathématique Sociales, Paris, France.


#### Abstract

It is generally believed that, when a linguistic item acquires a new meaning, its overall frequency of use in the language rises with time with an S-shaped growth curve. Yet, this claim has only been supported by a limited number of case studies. In this paper, we provide the first corpusbased quantitative confirmation of the genericity of the S-curve in language change. Moreover, we uncover another generic pattern, a latency phase of variable duration preceding the S-growth, during which the frequency of use of the semantically expanding word remains low and more or less constant. We also propose a usage-based model of language change supported by cognitive considerations, which predicts that both phases, the latency and the fast S-growth, take place. The driving mechanism is a stochastic dynamics, a random walk in the space of frequency of use. The underlying deterministic dynamics highlights the role of a control parameter, the strength of the cognitive impetus governing the onset of change, which tunes the system at the vicinity of a saddlenode bifurcation. In the neighborhood of the critical point, the latency phase corresponds to the diffusion time over the critical region, and the S-growth to the fast convergence that follows. The duration of the two phases is computed as specific first passage times of the random walk process, leading to distributions that fit well the ones extracted from our dataset. We argue that our results are not specific to the studied corpus, but apply to semantic change in general.


Language can be approached through three different, complementary perspectives. Ultimately, it exists in the mind of language users, so that it is a cognitive entity, rooted in a neuro-psychological basis. But language exists only because people interact with each other: It emerges as a convention among a community of speakers, and answers to their communicative needs. Thirdly, language can be seen as something in itself: An autonomous, emergent entity, obeying its own inner logic. If it was not for this third Dasein of language, it would be less obvious to speak of language change as such.

The social and cognitive nature of language informs and constrains this inner consistency. Zipf's law, for instance, may be seen as resulting from a trade-off between the ease of producing the utterance, and the ease of processing it [1]. It relies thus both on the cognitive grounding of the language, and on its communicative nature. Those two external facets of language, cognitive and sociological, are similarly expected to channel the regularities of linguistic change. Modeling attempts (see [2] for an overview) have explored both how sociolinguistic factors can shape the process of this change [3, 4] and how this change arises through language learning by new generations of users [5, 6]. Some models also consider mutations of language itself, without providing further details on the social or cognitive mechanisms of change [7]. In this paper, we propose to view language change as initiated by language use, which is the repeated call to one's linguistic resources in order to express oneself or to make sense of linguistic productions of others.

This approach is in line with exemplar models [8] and related works, such as the Utterance Selection Model [9] or the model proposed by Victorri [10, which describes an out-of-equilibrium shaping of semantic structure through repeated events of communication.

Leaving aside socio-linguistic factors, we focus on a cognitive approach of linguistic change, more precisely of semantic expansion. Semantic expansion occurs when a new meaning is gained by a word or a construction (we will henceforth refer more vaguely to a linguistic 'form', so as to remain as general as possible). For instance, way, in the construction way too, has come to serve as an intensifier (e.g. 'The only other newspaper in the history of Neopia is the Ugga Ugg Times, which, of course, is way too prehistoric to read.' [11]). The fact that polysemy is pervasive in any language [12] suggests that semantic expansion is a common process of language change and happens constantly throughout the history of a language. Grammaticalization [13] - a process by which forms acquire a (more) grammatical status, like the example of way too above - and other interesting phenomena of language change [14, 15], fall within the scope of semantic expansion.

Semantic change is known to be associated with an increase of frequency of the form whose meaning expands. This increase is expected indeed: As the form comes to carry more meanings, it is used in a broader number of contexts, hence more often. This implies that any instance of semantic change should have its empirical counterpart in the frequency rise of the use of the form. This
rise is furthermore believed to follow an S-curve [16, 17, yet such claim, to our knowledge, has not been quantitatively grounded on more than a few chosen examples. Besides, it is not easily accounted through theoretical modeling: In a sociolinguistic framework for instance, it requires either a very specific social structure, or the assumption that the new use is favored intrinsically [18]. Such a framework also suffers from what is known as the Threshold Problem, the fact that a novelty will fail to take over an entire community of speakers, because of the isolated status of an exceptional deviation [19].

In this paper, we provide a broad corpus-based investigation of the frequency patterns associated with a few hundred semantic expansions. It turns out that the Scurve pattern is corroborated, but must be completed by a preceding latency part, in which the frequency of the form does not significantly increase, even if the new meaning is already present in the language. To explain this surprising behavior, which seems to have escaped notice so far, we propose a usage-based model of the process of semantic expansion, implementing basic cognitive hypotheses regarding language use. By means of our model, we relate the micro-process of language use at the individual scale, to the observed macro-phenomenon of a recurring frequency pattern occurring in semantic expansion.

## I. QUANTIFICATION OF CHANGES IN A LARGE CORPUS

We worked on the French corpus Frantext [20], to our knowledge the only textual database allowing for a reliable study covering several centuries (see Material and Methods and Appendix A). We studied changes in frequency of use for 400 forms which have undergone one or several semantic expansions, on a time range going from 1321 up to nowadays. We choose forms so as to focus on semantic expansions leading to a functional meaning - such as discursive, prepositional, or procedural meanings. Semantic expansions whose outcome remains in the lexical realm (as the one undergone by sentence, whose meaning evolved from 'verdict, judgment' to 'meaningful string of words') have been left out. Functional meanings indeed present several advantages: They are often accompanied by a change of syntagmatic context, allowing to track the semantic expansion more accurately (e.g. way in way too + adj.); they are also less sensitive to socio-cultural and historical influences; finally they are less dependent on the specific content of a text, be it literary or academic.

The profiles of frequency of use extracted from the database are illustrated on Figure 1 for nine forms. We find that 286 cases display at least one sigmoidal increase of frequency in the course of their evolution, which makes up more than $70 \%$ of the total. We provide a small selection of the observed frequency patterns (Fig. 2a), whose associated logit transforms (Fig. 2b) follows a linear be-
havior, indicative of the sigmoidal nature of the growth (see Material and Methods). We thus find a robust statistical validation of the sigmoidal pattern, confirming the general claim made in the literature.

Furthermore, we find two major phenomena besides this sigmoidal pattern. The first one is that, in most cases, the final plateau towards which the frequency is expected to stabilize after its sigmoidal rise is not to be found: The frequency immediately starts to decrease after having reached a maximum (Fig. 11). However, such a decrease process is not symmetrical with the increase, in contrast with other cases of fashion-driven evolution in language, e.g. first names distribution [21]. Though this decrease may be, in a few handful of cases, imputable to the disappearance of a form (ex: après ce, replaced in Modern French by après quoi), in most cases it is more likely to be the sign of a narrowing of its uses.

The second feature is that the fast growth is very often preceded by a long latency up to several centuries, during which the new form is used, but with a comparatively low and rather stable frequency (Fig. 2a). One should note that the latency times may be underestimated: If the average frequency is very low during the latency part, the word may not show up at all in the corpus, especially in decades for which the available texts are sparse. The pattern of frequency increase is thus better conceived of as a latency followed by a growth, as exemplified by de toute façon (Fig. 3) - best translated by anyway in English, since the present meanings of these two terms are very close, and remarkably, despite quite different origins, the two have followed parallel paths of change.

To our knowledge, these two features, latency and absence of a stable plateau, have not been documented before, even though a number of specific cases of latency have been observed. For instance, it has been remarked in the case of just because that the fast increase is only one stage in the evolution [22]). In the following, we propose a model describing both the latency and the Sgrowth periods. We leave for future work the study of the decrease of frequency following the S-growth.

## II. A COGNITIVE SCENARIO

To account for the specific frequency pattern evidenced by our data analysis, we propose a scenario focusing on cognitive aspects of language use, leaving all sociolinguistic effects back-grounded by making use of a representative agent, mean-field type, approach. We limit ourselves to the case of a competition between two linguistic variants, given that most cases of semantic expansion can be understood as such, even if the two competing variants cannot always be explicitly identified. Initially, in some concept or context of use $C_{1}$, one of the two variants, henceforth noted $Y$, is systematically chosen, so that it conventionally expresses this concept. The question we address is thus how a new variant, say $X$, can be used in this context and eventually evict the old variant $Y$ ?


FIG. 1. Frequency evolution on the whole time range (1321-2020) of nine different forms. Each blue bar shows the frequency associated to a decade. Frequency has been multiplied by a $10^{5}$ factor for an easier reading.

## A. Hypotheses

The main hypothesis we propose is that the new variant almost never is a brand new merging of phonemes whose meaning would pop out of nowhere. As Haspelmath highlights [23], a new variant is almost always a periphrastic construction, i.e., actual parts of language, put together in a new, meaningful way. Furthermore, such a construction, though it may be exapted to a new use, may have showed up from time to time in the time course of the language history, in an entirely compositional way; this is the case for par ailleurs, which incidentally appears as early as the XIV ${ }^{\text {th }}$ in our corpus, but arises as a construction in its own right during the first part of the XIX ${ }^{\text {th }}$ century only. In other words, the use of a linguistic form $X$ in a context $C_{1}$ may be entirely new, but the form $X$ was most probably already there in another context of use $C_{0}$, or equivalently, with another meaning.

We make use of the well-grounded idea [24] that there exists links between concepts due to the intrinsic polysemy of language: There are no isolated meanings, as each concept is interwoven with many others, in a complicated tapestry. These links between concepts are asymmetrical, and they can express both universal mappings
between concepts [25, 26] and cultural ones (e.g. entrenched metaphors [27]). As the conceptual texture of language is a complex network of living relations rather than a collection of isolated and self-sufficient monads, semantic change is expected to happen as the natural course of language evolution and to occur repetitively throughout its history, so that at any point of time, there are always several parts of language which are undergoing changes. The simplest layout accounting for this network structure in a competitive situation consists then in two sites, such that one is influencing the other through a cognitive connexion of some sort.

## B. Model formalism

We now provide details on the modeling of a competition between two variants $X$ and $Y$ for a given context of use, or concept, $C_{1}$, also considering the effect exerted by the related context or concept $C_{0}$ on this evolution.

- Each concept $C_{i}, i=0,1$, is represented by a set of exemplars of the different linguistic forms. We note $N_{\mu}^{i}(t)$ the number at time $t$ of encoded exemplars (or occurrences) of form $\mu \in\{X, Y\}$, in context $C_{i}$, in the memory, of the representative agent.


FIG. 2. (a) A selection of frequency evolutions showing the latency period and the S-growth, separated by a red vertical line. (b) Logit transforms of the S-growth part of the preceding curves. Red dots correspond to data points and the green line to the linear fit of this set of points.

- The memory capacity of an individual being finite, the population of exemplars attached to each concept $C_{i}$ has a finite size $M_{i}$. For simplicity we assume that all memory sizes are equal $\left(M_{0}=M_{1}=M\right)$. As we consider only two forms $X$ and $Y$, for each $i$ the relation $N_{X}^{i}(t)+N_{Y}^{i}(t)=M$ always hold: We can focus on one of the two forms, here $X$, and drop out the form subscript, granted that all quantities refer to $X$.
- The absolute frequency $x_{t}^{i}$ of form $X$ at time $t$ in context $C_{i}$ - the fraction of 'balls' of type $X$ in the bag attached to $C_{i}$ - is thus given by the ratio $N^{i}(t) / M$. In the initial situation, $X$ and $Y$ are assumed to be established convention for respectively expressing $C_{0}$ and $C_{1}$, so that we start with $N^{0}(t=0)=M$ and $N^{1}(t=0)=0$.


FIG. 3. Overall evolution of the frequency of use of de toute façon (main panel), with focus on the S-shape increase (right inner panel), whose logit transformation follows a linear fit (left inner panel). Preceding the S-growth, one observes a long period of very low frequency (up to 34 decades).

- Finally, $C_{0}$ exerts an influence on context $C_{1}$, but this influence is assumed to be unilateral. Consequently, the content of $C_{0}$ will not change in the course of the evolution and we can focus on $C_{1}$. An absence of explicit indication of context is thus to be understood as referring to $C_{1}$.


## C. Dynamics

The dynamics of the system runs as follow. At each time $t$, one of the two linguistic forms is chosen to express concept $C_{1}$. The form $X$ is uttered with some probability $P(t)$, to be specified below, and $Y$ with probability $1-P(t)$. In order to keep constant the memory size of the population of occurrences in $C_{1}$, a past occurrence is randomly chosen (with a uniform distribution) and the new occurrence takes its place. This dynamics is then repeated a large number of times. Note that this model focuses on a speaker perspective (for alternative variants, see Appendix B).

We want to explicit the way $P(t)$ depends on $x(t)$, the absolute frequency of $X$ in this context at time $t$. The simplest choice would be $P(t)=x(t)$. However, we want to take into account several facts, as explained below.

- As context $C_{0}$ exerts an influence on context $C_{1}$, denoting by $\gamma$ the strength of this influence, we assume the probability $P$ to rather depend on an effective frequency $f(t)$ (Fig. 4a),

$$
\begin{equation*}
f(t)=\frac{N^{1}(t)+\gamma N^{0}(t)}{M+\gamma M}=\frac{x(t)+\gamma}{1+\gamma} \tag{1}
\end{equation*}
$$

- We now specify the probability $P(f)$ to select $X$ at time $t$ as a function of $f=f(t)$. First, $P(f)$ must be nonlinear. Otherwise, the change occurs with certainty as soon as the effective frequency $f$ of the novelty is nonzero, that is, insofar two meanings are related, the form expressing the former will also be recruited to express the
latter. This change would also start in too abrupt a way, while sudden, instantaneous takeovers are not known to happen in language change.

Second, one should preserve the symmetry between the two forms, that is, $P(f)=1-P(1-f)$, as well as verify $P(0)=0$ and $P(1)=1$. Note that this symmetry is stated in terms of the effective frequency $f$ instead of the actual frequency $x$, as production in one context always accounts for the contents of neighboring ones.

For the numerical simulations, we made the following specific choice which satisfies these constraints:

$$
\begin{equation*}
P(f)=\frac{1}{2}\left\{1+\tanh \left(\beta \frac{f-(1-f)}{\sqrt{f(1-f)}}\right)\right\} \tag{2}
\end{equation*}
$$

where $\beta$ is a parameter governing the non-linearity of the curve. Replacing $f$ in terms of $x$, the probability to choose $X$ is thus a function $P_{\gamma}(x)$ of the current absolute frequency $x$ :

$$
\begin{equation*}
P_{\gamma}(x)=\frac{1}{2}\left\{1+\tanh \left(\beta \frac{2 x-1+\gamma}{\sqrt{(x+\gamma)(1-x)}}\right)\right\} \tag{3}
\end{equation*}
$$

## D. Analysis: Bifurcation and latency time

The dynamics outlined above (Fig. 4b) is equivalent to a random walk on the segment $[0 ; 1]$ with a reflecting boundary at 0 and an absorbing one at 1 , and with steps of size $1 / M$. The probability of going forward at site $x$ is equal to $(1-x) P_{\gamma}(x)$, and the probability of going backward to $x\left(1-P_{\gamma}(x)\right)$.

For large $M$, a continuous, deterministic approximation of this random walk leads, after a rescaling of the time $M t \rightarrow t$, to the first order differential equation for $x(t)$ :

$$
\begin{equation*}
\dot{x}=P_{\gamma}(x)-x . \tag{4}
\end{equation*}
$$

This dynamics admits either one or three fixed points (Fig. 5a), $x=1$ always being one. Below a threshold value $\gamma_{c}$, which depends on the non-linearity parameter $\beta$, a saddle-node bifurcation occurs and two other fixed points appear. The system, starting from $x=0$, is stuck at the smallest stable fixed point. The transmission time, i.e. the time required for the system to go from 0 to 1 , is therefore infinite (Fig. 5b). Above the threshold value $\gamma_{c}$, only the fixed point $x=1$ remains, so that the new variant eventually takes over the context for which it is competing. Our model thus describes how the strengthening of a cognitive link can trigger a semantic expansion process.

Slightly above the transition, a stranglehold region appears where the speed almost vanishes. Accordingly, the time spent in this region diverges. The frequency of the new variant will stick to low values for a long time, in a way similar to the latent behavior evidenced by our dataset. This latency time in the process of change can


FIG. 4. (a) Difference between absolute frequency $x$ and relative frequency $f$ in context $C_{1}$. Absolute frequency $x$ is given by the ratio of $X$ occurrences encoded in $C_{1}$. Effective frequency $f$ also takes into account the $M$ occurrences contained in the influential context $C_{0}$, with a weight $\gamma$ standing for the strength of this influence. (b) Schematic view of the process. At each iteration, either $X$ or $Y$ is chosen to be produced and thus encoded in memory, with respective probability $P_{\gamma}(x)$ and $1-P_{\gamma}(x)$; the produced occurrence is here represented in the purple capsule. Another occurrence, already encoded in the memory, is uniformly chosen to be erased (red circle) so as to keep the population size constant. Hence the number of $X$ occurrences, $N_{X}$, either increases by 1 if $X$ is produced and $Y$ erased, decreases by 1 if $Y$ is produced and $X$ erased, or remains constant if the erased occurrence is the same as the one produced.
thus be understood as a near-critical slowing down of the underlying dynamics.

Past this deterministic approximation, there is no more clear-cut transition (Fig. 5b) and the above explanation needs to be refined. The deterministic speed can be understood as a drift velocity of the Brownian motion on the $[0 ; 1]$ segment, so that in the region where the speed vanishes, the system does not move in average. In this region of vanishing drift, the frequency fluctuates over a small set of values and does not evolve significantly over time. Once it escapes this region, the drift velocity drives the process again, and the replacement process takes off. Latency time can thus be understood as a first-passage time out of a trapping region.

## III. NUMERICAL RESULTS

## A. Model simulations

We ran numerical simulations of the process described above (Fig. 4b), with the following choice of parameters: $\beta=0.808, \delta=0.0$ and $M=5000$, where $\delta=\left(\gamma-\gamma_{c}\right) / \gamma_{c}$ is the distance to the threshold. The specific value of $\beta$ corresponds to a maximization of $x_{c}$, the frequency value at which the system gets stuck. It reflects the assumption that the linguistic system should allow for synonymic variation in the situation where no replacement takes place. We chose $\delta=0.0$ in order for the system to be purely diffusive in the vicinity of $x_{c}$. The choice of $M$


FIG. 5. (a) Speed $\dot{x}$ of the deterministic process for each of the sites, for different values of $\beta$ and $\delta=\left(\gamma-\gamma_{c}\right) / \gamma_{c}$, the distance to threshold. Depending on the sign of $\delta$, there is either one or three fixed points. (b) Inverse transmission time (time required for the system to go from 0 to 1 ), for the deterministic process (blue dotted line), and for the averaged stochastic process (green line), as a function of the control parameter $\delta$. Deterministic transmission time diverges at the transition while averaged stochastic transmission time remains finite.
is arbitrary.
From the model simulations, data is extracted and analyzed in two parallel ways. On one side, simulations provide surrogate data: We can mimic the corpus data analysis and count how many tokens of the new variant are produced in a given timespan (set equal to $M$ ), to be compared with the total number of tokens produced in this timespan. We then extract 'empirical' latency and growing times (Fig. 6a), applying the same procedure as for the corpus data.

One the other side, for each run we track down the position of the walker, which is the frequency $x(t)$ achieved by the new variant at time $t$. This allows to compute first passage times. We then alternatively compute analytical latency and growth times ('analytical' to distinguish them from the former 'empirical' times) as follows. Latency time is here defined as the difference between the first-passage times at the exit and the entrance of a 'trap' region (see Appendix $\mathbb{C}$ for additional details). Analytical growth time is defined as the remaining time of the process once this exit has been reached. Their distribution over 10,000 runs of the process are fitted with Inverse Gaussian distribution, which would be the expected distributions if the jump probabilities were homogeneous over the corresponding regions (an approximation then better suited for latency time than for growth time). Figure 6d shows the remarkable agreement between the 'empirical' and 'analytical' approaches, together with the quality of the fits with the Inverse Gaussian distribution.

Crucially, those two macroscopic phenomena, latency and growth, are thus to be understood as of the same nature, which explains why their statistical distribution must be of the same kind. Furthermore, the boundaries of the trap region leading to the best correspondence between first passage times and empirically determined latency and growth times are meaningful, as they correspond to the region where the uncertainty on the transmission time significantly decreases (Fig. 6b).

## B. Confrontation with corpus data

Our model predicts that both latency and growth times should be governed by the same kind of statistics, Inverse Gaussian being a suited approximation of those. Inverse Gaussian distribution is governed by two parameters, its mean $\mu$ and a parameter $\lambda$ given by the ratio $\mu^{3} / \sigma^{2}$, $\sigma^{2}$ being the variance. We fit the empirical histograms with an Inverse Gaussian distribution whose parameters are given by the empirical mean and variance of the relevant quantities. We find a good agreement for both the latency and the growth times (Fig. 77).

Although there are short growth times in the frequency patterns of the forms we studied, below six decades they are not described by enough data points to assess reliably the specificity of the sigmoid fit. On the histogram there is therefore no data for these growth times. This issue is further discussed in Appendix D. However, the distribution must decrease when growth time approaches 0 (notably an exponential fit is to be ruled out); otherwise, instantaneous changes would be far too numerous, so that language would be completely unstable. The decrease predicted by the Inverse Gaussian is realistic in this aspect.

The main quantitative features extracted from the dataset are thus correctly mirrored by the behavior of our model. We confronted the model with the data on other quantities, such as the correlation between growth time and latency time. There again, the model proves to match appropriately quantitative aspects of semantic expansion processes Appendix E.

## IV. DISCUSSION

Based on a corpus-based analysis of frequency of use, we have uncovered two robust stylized facts of semantic change: an S-curve of frequency growth, preceded by a latency period where the semantic change has already taken place while the frequency remains low. We have proposed a model predicting that these two features, albeit qualitatively quite different, are two aspects of one and the same phenomenon.

The hypotheses on which this model lies are wellgrounded on claims from Cognitive Linguistics: Language is resilient to change (non-linearity of the $P$ function); language users have cognitive limitations; the semantic territory is organized as a network whose neighboring sites are asymmetrically influencing each other. The overall agreement with empirical data tends to suggest that language change may indeed be cognitively driven by semantic bridges of different kinds between the concepts of the mind, and constrained by the mnemonic limitations of this very same mind. We note that our model may however be given a different, purely sociolinguistic, interpretation: this, together with the limits of such a view point, is discussed in Appendix B 5

According to our model, the onset of change depends


FIG. 6. (a) Time evolution of the frequency of produced occurrences (output of a single run). Growth part and latency part are shown respectively in blue and red. The logit transform (with linear fit) of the growth is shown in the inset. (b) Distribution of latency time (top) and growth time (bottom) over $10 k$ processes, extracted from an empirical approach (blue wide histogram) and a first-passage time one (magenta thin histogram), with their respective Inverse Gaussian fits (in red: Empirical approach; in green: First-passage time approach). (c) Uncertainty on the transmission time given the position of the walker. The entrance and the exit of the trap are shown, respectively, by green and magenta line. The trap corresponds to the region where the uncertainty drops from a high value to a low value.
on the strength of the conceptual link between the source context and the target context: If the link is strong enough, that is, above a given threshold, it serves as a channel so that a form can 'invade' the target context and then oust the previously established form. In a sense, the sole existence of this cognitive mapping is already a semantic expansion of some sort, yet not necessarily translated into linguistic use. Latency is specifically understood as resulting from a near-critical behavior: If the link is barely strong enough for the change to take off, then the channel becomes extremely tight and the invasion process slows down drastically. These narrow channels are likely to be found between lexical and grammatical meanings [28, 29]. This would explain why
the latency-growth pattern is much more prominent in the processes of grammaticalization, positing latency as a phenomenological hint of this latter category.

Finally, we argue that our results, though grounded on instances of semantic expansion in French, apply to semantic expansion in general. The time period covered is long enough ( 700 years) to exclude the possibility that our results be ascribable to a specific historical, sociological, or cultural context. The French language itself has evolved, so that Middle French and contemporary French could be considered as two different languages, yet our analysis apply to both indistinctly. Besides, the latencygrowth pattern is to be found in other languages; for instance, Google Ngram queries for constructions such


FIG. 7. Inversian Gaussian fit of the latency times (left) and the growth times (right) extracted from corpus data. Parameters are computed from the mean and the variance of the data. Data points are shown by a blue histogram, the Inverse Gaussian fit being represented as red dots. The discrepancy observed for six decades is discussed in Appendix D
as way too, save for, no matter what, yield qualitative frequency profiles consistent with our claims. Our model also tends to confirm the genericity of this pattern, as it relies on cognitive mechanisms whose universality has been well evidenced 30.

## V. MATERIALS AND METHODS

We worked on the Frantext corpus [20], which in 2016 contained for the chosen time range 4674 texts and 232 millions of words. More details are given in Appendix A. It would have been tempting to make use of the large database Google Ngram, yet it was not deemed appropriate for out study, as we explain in Appendix F.

We studied changes in frequency of use for nearly 400 instances of semantic expansion processes in French, on a time range going from 1321 up to nowadays. See Appendix $G$ for a complete list of the studied forms.

## A. Extracting patterns from corpus data

a. Measuring frequencies. We divided our corpus into 70 decades. Then, for each form, we recorded the number of occurrences per decade, dividing this number by the total number of occurrences in the database for that decade. The output number is called here the frequency of the occurrence for the decade, and is noted $x_{i}$ for decade $i$. In order to smooth the obtained data, we replaced $x_{i}$ by a moving average, taht is, for $i \geq i_{0}+4, i_{0}$ being the first decade of our corpus: $x_{i} \leftarrow \frac{1}{5} \sum_{k=i-4}^{i} x_{k}$.
b. Sigmoids. We looked for major increases of frequency. When such a major shift is encountered, we automatically (see below) identify frequencies $x_{\text {min }}$ and $x_{\text {max }}$, respectively at the beginning and the end of the
increasing period. If we respectively note $i_{\text {start }}$ and $i_{\text {end }}$ the decades for which $x_{\text {min }}$ and $x_{\text {max }}$ are reached, then we define the duration $w$ of the increasing period as $w=i_{\text {end }}-i_{\text {start }}+1$. To quantify the sigmoidal nature of this growth pattern, we apply the logit transformation to the frequency between $x_{\min }$ and $x_{\max }$ :

$$
\begin{equation*}
y_{i}=\log \left(\frac{x_{i}-x_{\min }}{x_{\max }-x_{i}}\right) \tag{5}
\end{equation*}
$$

If the process follows a sigmoid $\tilde{x}_{i}$ of equation:

$$
\begin{equation*}
\tilde{x}_{i}=x_{\min }+\frac{x_{\max }}{1+e^{-h i-b}} \tag{6}
\end{equation*}
$$

then the logit transform of this sigmoid satisfies: $\tilde{y}_{i}=$ $h i+b$. We thus fit the $y_{i}$ 's given by (5) with a linear function, which gives the slope $h$ associated with it, the residual $r^{2}$ quantifying the quality of the fit. The boundaries $i_{\text {start }}$ and $i_{\text {end }}$ have been chosen so as to maximize $w$, with the constraint that the $r^{2}$ of the linear fit should be at least equal to a value depending on the number of the data points.
c. Latency period. In most cases ( $74 \%$ of sigmoidal growths), one observes that the fast increasing part is preceded by a phase during which the frequency remains constant or nearly constant. The duration of this part, denoted by $T_{1}$ in this paper, is identified automatically as follows. Starting from the decade $i_{\text {start }}$, previous decades $j$ are included in the latency period as long as they verify $\left|x_{j}-x_{\text {min }}\right|<0.15 *\left(x_{\max }-x_{\min }\right)$ and $x_{j}>0$, and cease to be included either as soon as the first condition is not verified, or if the second condition does not hold for a period longer than 5 decades. Then the start $i_{l a t}$ of the latency point is defined as the lowest $j$ verifying both conditions, so that $T_{1}$ is given by $T_{1}=i_{\text {start }}-i_{\text {lat }}$.

## ACKNOWLEDGMENTS

We thank B. Derrida for a useful discussion on random walks. QF acknowledges a fellowship from PSL Research

University. BF is a CNRS member. JPN is senior researcher at CNRS and director of studies at the EHESS.
[1] Ramon Ferrer i Cancho and Ricard V Solé. Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences, 100(3):788791, 2003.
[2] Quentin Feltgen, Benjamin Fagard, and Jean-Pierre Nadal. Modeling Language Change: The Pitfall of Grammaticalization, pages 49-72. Springer, 2017.
[3] Vittorio Loreto, Andrea Baronchelli, Animesh Mukherjee, Andrea Puglisi, and Francesca Tria. Statistical physics of language dynamics. Journal of Statistical Mechanics: Theory and Experiment, 2011(04):P04006, 2011.
[4] Jinyun Ke, Tao Gong, and William SY Wang. Language change and social networks. Communications in Computational Physics, 3(4):935-949, 2008.
[5] Martin A Nowak, Natalia L Komarova, and Partha Niyogi. Computational and evolutionary aspects of language. Nature, 417(6889):611-617, 2002.
[6] Thomas L Griffiths and Michael L Kalish. Language evolution by iterated learning with Bayesian agents. Cognitive Science, 31(3):441-480, 2007.
[7] Igor Yanovich. Genetic Drift Explains Sapir's "drift" In Semantic Change. In S.G. Roberts, C. Cuskley, L. McCrohon, O. Barceló-Coblijn, and T. Verhoef, editors, The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11), pages 321-329, 2016.
[8] Janet B Pierrehumbert. Exemplar dynamiccs: Word frequency, lenition and contrast, volume 45, page 137. John Benjamins Publishing, 2001.
[9] Gareth J Baxter, Richard A Blythe, William Croft, and Alan J McKane. Utterance selection model of language change. Physical Review E, 73(4):046118, 2006.
[10] Bernard Victorri. The use of continuity in modeling semantic phenomena. In Fuchs, C. and Victorri, B., editor, Continuity in linguistic semantics, pages 241-251. Benjamins, 1994.
[11] Neopets Inc. The Neopian Times. http://www.neopets. com/ntimes/index.phtml?section=497377\&issue=457, 2010.
[12] Sabine Ploux, Armelle Boussidan, and Hyungsuk Ji. The semantic atlas: an interactive model of lexical representation. In Proceedings of the seventh conference of International Language Resources ans Evaluation, pages 1-5, 2010.
[13] Paul J Hopper and Elizabeth Closs Traugott. Grammaticalization. Cambridge University Press, 2003.
[14] Britt Erman and U-B Kotsinas. Pragmaticalization: the case of ba'and you know. Studier $i$ modern språkvetenskap, 10:76-93, 1993.
[15] Laurel J Brinton and Elizabeth Closs Traugott. Lexicalization and language change. Cambridge Univ. Press, 2005.
[16] Anthony Kroch. Reflexes of grammar in patterns of language change. Language variation and change, 1(3):199244, 1989.
[17] Jean Aitchison. Language change: progress or decay? Cambridge University Press, 2013.
[18] Richard A Blythe and William Croft. S-curves and the mechanisms of propagation in language change. Language, 88(2):269-304, 2012.
[19] Daniel Nettle. Using social impact theory to simulate language change. Lingua, 108(2):95-117, 1999.
[20] ATILF. FRANTEXT textual database, http://www.frantext.fr., Octobre 2014.
[21] Baptiste Coulmont, Virginie Supervie, and Romulus Breban. The diffusion dynamics of choice: From durable goods markets to fashion first names. Complexity, 2015.
[22] Martin Hilpert and Stefan Th Gries. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. Literary and Linguistic Computing, 24(4): 385-401, 2009.
[23] Martin Haspelmath. Why is grammaticalization irreversible? Linguistics, 37(6):1043-1068, 1999.
[24] Richard Hudson. Language networks: the new Word Grammar. Oxford University Press, 2007.
[25] Bernd Heine. Cognitive foundations of grammar. Oxford University Press, 1997.
[26] Johannes Dellert. Using Causal Inference To Detect Directional Tendencies In Semantic Evolution. In S.G. Roberts, C. Cuskley, L. McCrohon, O. BarcelóCoblijn, and T. Verhoef, editors, The Evolution of Language: Proc. of the 11th International Conference (EVOLANG11), pages 88-96, 2016.
[27] George Lakoff and Mark Johnson. Metaphors we live by. University of Chicago press, 1980.
[28] Bernd Heine. On the role of context in grammaticalization, volume 49, pages 83-102. 2002.
[29] Gabriele Diewald. Context types in grammaticalization as constructions. Constructions, 1(9), 2006.
[30] Bernd Heine and Tania Kuteva. World lexicon of grammaticalization. Cambridge University Press, 2002.
[31] Christiane Marchello-Nizia. L'oral représenté: un accès construit à une face cachée des langues 'mortes', pages 247-264. Peter Lang, 2012.
[32] Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. Verbs change more than nouns: a bottom-up computational approach to semantic change. Lingue e linguaggio, 15(1):7-28, 2016.
[33] Sara Graça Da Silva and Jamshid J Tehrani. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. Royal Society open science, 3 (1):150645, 2016.
[34] Sidney Redner. A guide to first-passage processes. Cambridge University Press, 2001.
[35] Bruno Gaume, Karine Duvignau, and Martine Vanhove. Semantic associations and confluences in paradigmatic networks. From Polysemy to Semantic Change Towards a typology of lexical semantic associations, John Benjamins, pages 233-264, 2008.
[36] Quentin Michard and J-P Bouchaud. Theory of collective opinion shifts: from smooth trends to abrupt swings. The European Physical Journal B-Condensed Matter and Complex Systems, 47(1):151-159, 2005.
[37] Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. PloS one, 10(10):e0137041, 2015.

## Appendix A: Textual data base Frantext

The data we collected for the present study comes from the Frantext database [20, one of the most extensive databases available in French, to which one has access under subscription by the ATILF-CNRS laboratory. Frantext is an ever-expanding gathering of 4,746 texts to this day (8th december 2016), updated every year. This corpus presents various literary genres (epistolary, drama, poetry, essays, scientific books), but mainly novels, almost exclusively from French literature (with a few translated works). The publication year of the texts range from 950 to 2013 . The allotment of the texts between the different time periods is however far from being homogeneous, and most of them belong to the twentieth century: Indeed, the number of texts by decade roughly follows an exponential increase (Fig. 8).

Frantext, while being much smaller than Google Ngram, provides much cleaner and more controlled results (see E). We decided to start from the decade 13211330, as from this date all decades are associated with at least seven texts. In our corpus, we retained most of the texts, with a few exceptions, e.g. when the date provided by Frantext was unsatisfying (for instance, the text referred to as 6205, Le Canarien, pièces justificatives is dated 'between 1327 and 1470'), or when we knew that the text has been written over too long a time period, as is the case for the text Chartes et documents de l'abbaye de Saint-Magloire (ref 8203), whose publication year (1330) is far from covering the time span during which the document was compiled. Most interestingly, Frantext also provides the surrounding text on which a token is to be found, so that it is possible to check if the different occurrences make sense and truly correspond to the request.

Frantext is not flawless. Some parts of the scanned texts have been appended through posterior editing. This is clearly the case for the text A017, Chroniques de Morée, where some page notes from a contemporaneous edition of this medieval chronicle have been included, so that the request for 'dans' may return an occurrence such as 'Erreur dans la numérotation de l'édition' ('error in the edition numbering'). Some decades are also strongly unbalanced in the available texts. For instance, among the 2.7 million words of decade 1551-1560, more than one third of them comes from the works of a single author, Jean Calvin (references E198, B022, R849 to R852). Another bias comes from the fact that drama pieces, up to the end of the Modern Era, were making use of represented orality 31] much more than literary texts, so that many new constructions appear in them before spreading among the other texts. This would not be a problem if the proportion of dramas were more or less constant across the decades, which is not the case. This problem vanishes in more recent times, when represented orality appears also frequently in novels, while drama becomes itself more sophisticated and shifts further away from daily language.

Frantext is not only a database. It comes also with built-in text-mining algorithms which allow to submit very refined queries to the database. Such queries can make use of booleans and a given number of blank words. For instance, the query (à $\mid \mathbf{a}) \& \mathbf{q}(\mathbf{1 , 2})$ (insçu|insu|insceu) ( $\& \mathbf{q}(\mathbf{1 , 2 )}$ is a blank slot for any one or two words) will retrieve occurrences such as $\grave{a}$ l'insu, à leur insu, but also à son propre insu. This kind of flexible requests are especially relevant when one is looking for specific constructions with a filling slot, as the corresponding possibilities cannot be exhaustively predicted. We studied for instance the construction d'une voix + ADJ. If we cannot list all adjectives, we can rule out all the parasite occurrences with an elaborated request such as ^(tous|receus) d'une voix ${ }^{\wedge}$ (que|qui|qu’|et|ensemble|trestous|de|d'|vous|le|la| les|par|pour|dont|-|.|;|,|:), where ^ and | respectively stands for the booleans 'not' and 'or'. Such a request makes it possible to capture unexpected adjectival constructs such as toute changée, si peu effroyée or extraordinairement rauque et rouillée, while discarding all spurious occurrences. Frantext also allows for special requests, for instance if one wishes to encompass several orthographic variations in a single query, for instance souventes?f* captures all possible variants of souventesfois, such as souventeffoiz, souvente fois, souventez fois, souventefoys, etc. This kind of elaborations prove to be all the more useful in the first stages of the evolution, where a functional construction has not yet become entrenched into an idiomatic form and can still be found in a high diversity of variants.

Once a request is submitted to the database, Frantext returns a datafile whose contents may vary according to the needs of the user. Depending on the options one chooses, the file displays, for each text, the text reference, the publication year, and the total number of occurrences of the query in that text. Next to this automatized procedure, we can also look across all individual occurrences in their context, as a sanity check. This was used frequently to help refining our queries. Unfortunately, it was impossible to ask Frantext for a file providing the statistics of the corpus itself, listing the number of occurrences per text reference. We extracted this information from an HTML page which does display this information (Corpus de travail $>$ Visualiser). The data file provided by Frantext was then directly treated by our own algorithm to compute average frequencies for each decade.

## A note on French

We acknowledge that we restricted ourselves to instances of semantic expansions in French, a choice which may appear to restrict the scope of our findings. As we argue in the main text, we believe this is not the case. In the following, we stress, 1 - the necessity to conduct the analysis on a long timescale (i.e. long enough so that we can consider the language to have changed dur-


FIG. 8. Number of millions of occurrences per decade in the Frantext database. Exponential fit is shown by a red line.
ing that period, just as contemporary French has drifted sufficiently away from Middle French (XIV ${ }^{\text {th }}$ century) so that, without specific training, the latter is only partially intelligible to speakers of the former), 2 - that few corpora are as efficient as Frantext to achieve such a goal.

Given the issues addressed in this paper, it appears important to consider instances taken from a large time period (seven centuries in our case). Indeed, a frequently asked question is whether or not recent technological advances (radio, TV, the Internet) have had an influence on the way language changes. Sociologically, this influence is obvious: Languages tend to homogenize over greater geographical areas and dialects have constantly declined throughout the twentieth century. Yet, the pattern of change of an established language is something entirely different. Our statistical survey shows that the pattern of change is the same, no matter in which century it may happen. It is furthermore consistent with recent findings establishing that the rate of change did not increase in the most recent decades [32]. It also goes along our claim that the pattern we exhibit is cognitively driven by memory retrieval and conceptual organization, two cognitive mechanisms that the most recent technological evolutions could not have significantly altered.

Alas, finding appropriate corpora covering a long time period in a given language is not obvious. As discussed in this SI, section S5, Google Ngram cannot be used for texts earlier than the nineteenth centuries, since the scanning procedure does not lead to reliable digital data. For the English language, the reputed British National Corpus restricts itself to the twentieth century. The Helsinki Corpus spans a time period suited for our purposes, but the texts are too sparse ( 450 in total) for the corpus to be fitted for a statistical survey. The CORDE
corpus, in Spanish, spans several centuries (XIII ${ }^{\text {th }}$ to $\mathrm{XX}^{\text {th }}$ ), and gathers an impressive amount of data as well ( 250 M words), but it covers different variants of Spanish (Argentinian, Colombian, Castillan, etc.) which cannot be blended together when it comes to investigate semantic expansions (note that CORDE dutifully offers to treat them apart, but then the database is not extensive enough for each of the variant separately). The querying system also suffers from serious limitations, and it is not possible to submit complex queries as is the case with Frantext. This latter database is therefore truly remarkable in many aspects and has to be considered an exception. We thus leave to further studies the case of other languages.

A last remark is in order: We deliberately do not provide any translation of the studied forms (SI, Table S1), however obscure they may appear to the reader. Indeed, these forms have all undergone a semantic expansion, so that a translation would be most mistaking as it would concern only one among several meanings adopted by the form. The only satisfying way of glossing the items we studied would have been to find forms which not only have the same meaning, but have also undergone (at least roughly) the same meaning shifts, as in the case of anyway and de toute façon for the later stages of their respective semantic evolutions. Obviously, this would have been possible only for a handful of cases, and we chose to leave the items without translation.

## Appendix B: Model variants

The model we propose in the main body of the paper describes a mechanism associated with language produc-
tion: It is solely based on a speaker perspective. Yet, language change may not come only from innovation in producing language, but also in understanding it. Actually, these two aspects cannot be separated: If an innovation is possible in a speaker perspective, it must also be accessible from a hearer perspective. Be it a speaker or a hearer, a language user relies on the same cognitive entity. It seems thus necessary to consider model variants where the novelty can come from this complementary perspective, as well as from a combination of the two.

## 1. Hearer variant

Let us consider the same situation as for the listener model: There are two meanings, $C_{0}$ and $C_{1}$, to which are attached a pool of memories of linguistic tokens. Initially, $C_{0}$ is populated by $X$ tokens only, while $C_{1}$ is populated by $Y$ tokens only. Just as context $C_{1}$ is fed by the memory of $C_{0}$ when it came to express $C_{1}$, if a linguistic occurrence yields meaning $C_{0}$, it can elicit meaning $C_{1}$ as well. Occurrences of $X$ thus have a chance to populate context $C_{1}$, so that we will note $x$ the proportion of $X$ tokens in $C_{1}$, just as we did in the speaker-based model. If we ascribe to the inference $C_{0} \Rightarrow C_{1}$ a probability equal to $\gamma$, then we can describe the dynamics as follows:

1. Either $C_{0}$ or $C_{1}$ are chosen to be expressed, with equal probabilities.
2. If $C_{0}$ has been chosen, $X$ is produced. If $C_{1}$ has been chosen, $X$ is produced with probability $P_{0}(x)$, otherwise $Y$ is produced. $P_{0}(x)$ is the same function as $P_{\gamma}(x)$, except that $\gamma$ is now set to 0 (there is no such thing as an effective frequency in this framework).
3. The produced occurrence is recorded in the chosen context. If $C_{0}$ has been chosen, an additional occurrence of the same kind as the previous one is recorded in $C_{1}$ with probability $\gamma\left(C_{0}\right.$ has elicited the meaning $C_{1}$ ).
4. A past occurrence is deleted whenever needed, so as to keep both memory sizes constant.

These dynamics correspond once more to a random walk where the jump probabilities, forward and backward, respectively $R^{H}(x)$ and $L^{H}(x)$ (where $H$ stand for 'hearer'), are given by:

$$
\begin{cases}R^{H}(x) & =\frac{1}{2}\left[\gamma+P_{0}(x)\right](1-x)  \tag{B1}\\ L^{H}(x) & =\frac{1}{2}\left(1-P_{0}(x)\right) x\end{cases}
$$

to be compared with the jump probabilities in the speaker perspective (respectively $L^{S}(x)$ and $R^{S}(x)$ for
the forward and backward jump probabilities):

$$
\left\{\begin{array}{l}
R^{S}(x)=P_{\gamma}(x)(1-x)  \tag{B2}\\
L^{S}(x)=\left(1-P_{\gamma}(x)\right) x
\end{array}\right.
$$

These modified jump probabilities lead to a new expression for the drift velocity:

$$
\begin{equation*}
\dot{x}=\frac{1}{2}\left[P_{0}(x)-x+\gamma(1-x)\right] . \tag{B3}
\end{equation*}
$$

A change of variable $y=(1+\gamma) x-\gamma$ leads to the same equation as equation 5 of the main paper, with a slightly different timescale accounting for the fact that two contexts are now being called:

$$
\begin{equation*}
\frac{2}{1+\gamma} \dot{y}=\left[P_{0}\left(\frac{y+\gamma}{1+\gamma}\right)-y\right] . \tag{B4}
\end{equation*}
$$

Indeed, $P_{0}\left(\frac{y+\gamma}{1+\gamma}\right)$ is exactly $P_{\gamma}(y)$, so that the fixed point in the hearer perspective $x_{c}^{H}$ will be given, as a function of the fixed point $x_{c}^{S}$ of the speaker perspective, as:

$$
\begin{equation*}
x_{c}^{H}=\frac{x_{c}^{S}+\gamma}{1+\gamma} \tag{B5}
\end{equation*}
$$

which is higher than $x_{c}^{S}$. This means that, in the hearer pespective, the latency frequency will also be higher. However, it does not entail that the change will be more or less likely to happen, since what triggers the change is the fact that $\gamma$ is equal to $\gamma_{c}$ or higher, and this parameter $\gamma_{c}$ remains the same throughout the perspective shift.

## 2. Combined model

We can now combine the Listener and Hearer perspectives, by taking into account the effective frequency $f$ instead of the actual frequency $x$ in step 2 of the dynamics outlined in the previous subsection. Then, in the above formulae, all $P_{0}(x)$ become $P_{\gamma}(x)$ (or equivalently, $\left.P_{0}(f)\right)$. The velocity is now set to:

$$
\begin{equation*}
\dot{x}=\frac{1}{2}\left[P_{\gamma}(x)-x+\gamma(1-x)\right] \tag{B6}
\end{equation*}
$$

Setting $X=(x+\gamma) /(1+\gamma)$, we get:

$$
\begin{equation*}
2(1+\gamma) \dot{x}=P_{0}(X)-X+(1-X) \gamma(2+\gamma) \tag{B7}
\end{equation*}
$$

We can now define a renormalized parameter $\tilde{\gamma}=\gamma(2+\gamma)$ to make this velocity similar to the one given by $(\overline{\mathrm{B}} 3)$. Setting $Y=(1+\tilde{\gamma}) X-\tilde{\gamma}$, we finally get:

$$
\begin{equation*}
2 \frac{1+\gamma}{1+\tilde{\gamma}} \dot{Y}=P_{\tilde{\gamma}}(Y)-Y \tag{B8}
\end{equation*}
$$

This implies that $\left(Y_{C}, \tilde{\gamma}_{c}\right)=\left(x_{c}^{S}, \gamma_{c}^{S}\right)$, so that the critical point $\left(x_{c}^{T}, \gamma_{c}^{T}\right)$ in this combined perspective is equal to:

$$
\begin{equation*}
\left(x_{c}^{T}, \gamma_{c}^{T}\right)=\left(\frac{x_{c}^{S}+\gamma_{c}^{S}}{1+\gamma_{c}^{S}}, \sqrt{1+\gamma_{c}^{S}}-1\right) \tag{B9}
\end{equation*}
$$

In this case $\gamma_{c}^{T}$ is lower than its hearer and speaker perspectives counterparts. It entails that the change would happen more easily. $x_{c}^{T}$ is somewhere in between $x_{c}^{S}$ and $x_{c}^{H}$.

## 3. Summary

All three variants of the model give rise to the same picture of sigmoidal growth preceded by a period of latency. The data does not allow to discriminate between either one of these three possibilities. Yet, the hypothesis that the change is driven by both hearer and speaker mechanisms is the most probable, as all language users adopt the role of hearer and speaker alternatively. An enthralling perspective of research would be to devise a quantitative criterion so as to see which of the three mechanisms best account for real language data. One could also investigate which features of language change speaker and hearer perspectives are respectively able to account for independently, and if some features need the conjunction of both to appear. Obviously, all those questions hinge upon available data and the finding of relevant observable quantities to look at.

## 4. Interpretations of the cognitive strength $\gamma$

In the proposed model, we make the assumption that all memory sizes are equal in the speaker perspective, and that all meanings $C_{i}$ are expressed with equal probability in the hearer perspective. Here we consider the alternative that the links in the network are not weighted: They are either 1 or 0 . The asymmetric structure between the two contexts $C_{0}$ and $C_{1}$ is however maintained.

## a. Heterogeneous memory sizes

Now let us assume different memory sizes for the two concepts, denoting by $m$ and $M$ the memory sizes of $C_{0}$ and $C_{1}$, respectively. Then the effective frequency of $X$ in $C_{1}$ is given by:

$$
\begin{equation*}
f=\frac{N+m}{M+m}=\frac{x+m / M}{1+m / M} \tag{B10}
\end{equation*}
$$

By defining $\gamma$ as the ratio of memories $m / M$, we recover the same effective frequency as before.

This means that the strength $\gamma$ of the cognitive link can be interpreted as a ratio between memory sizes. If all sites were connected to each other, the occurrences
expressing the contexts whose associated memory is the greatest would spread all over the network. However, not all sites lead to all others: There are pathways in the conceptual organization, which constrain possible semantic changes and allow for low-memory contexts to invade higher-memory ones.

The main difference brought forth by this interpretation is that it allows for $\gamma$ 's greater than one. In general, there would be no critical behavior and thus no latency, except if the conquering occurrence type comes from a very low memory context. This would suggest that, as grammaticalizations are well-characterized by the latency-growth pattern with sigmoidal increase, lexical meanings are allocated a much smaller memory than grammatical ones. However, it would also be the case within the lexicon, when a word goes from a concrete meaning to an abstract one.

It is not clear why functional and abstract meanings should be allocated a greater memory than concrete meanings. There could be for instance some advantage in making the more abstract and structural part of the conceptual realm more stable in their linguistic expression than other parts of speech, especially because they serve to constrain the processing of utterances and provide structure to the flow of speech. Were it the case, then we could understand the strong asymmetry evidenced by grammaticalization - the fact that lexical forms are recruited to express grammatical meanings overwhelmingly more frequently than the reverse. Indeed, if the links were from the stable (i.e. supported by a large memory size) to the unstable parts of the language, then all those links would be associated to a very high $\gamma$ parameter, so that all parts of language would soon come to be expressed by the grammatical forms. This would right away lead to a complete communicative failure. There would thus be an obvious advantage in preventing the links from grammatical concepts to lexical ones, hence in the unidirectionality exhibited by grammaticalization.

## b. Different probabilities of use

We now introduce different calling probabilities for $C_{0}$ and $C_{1}$ in the hearer perspective. Let's say that the probability to call $C_{0}$ is $\alpha$. Here again $\gamma$ is set to 1 (i.e. $C_{0}$ automatically entails $C_{1}$ ). The jump probabilities becomes thus:

$$
\begin{equation*}
R^{H}(x)=\left[\alpha+(1-\alpha) P_{0}(x)\right](1-x) \tag{B11}
\end{equation*}
$$

and:

$$
\begin{equation*}
L^{H}(x)=(1-\alpha)\left(1-P_{0}(x)\right) x \tag{B12}
\end{equation*}
$$

We can factorize $R^{H}(x)$ by $1-\alpha$. Then we recover the same computation as before, with the ratio of calling probabilities $\alpha /(1-\alpha)$ playing the role of $\gamma$. Furthermore, if we set the call probability to be proportional to memory size, then we recover the same $\gamma$ as in the preceding subsection. This assumption seems natural, since
greater memory sizes would help stabilizing the linguistic expressions of widely used meanings.

In such case, the near-criticality associated to the latency-growth pattern is recovered only if the links in the conceptual network are from the seldom called contexts to the often called contexts (so as to insure low enough values of $\gamma$ ). This seems a natural assumption for grammaticalization phenomena, since functional meanings are much more frequently called than lexical ones. Such assumption remains of course to be carefully investigated.

These two interpretations of the cognitive link point in the same direction: In short, the links of the conceptual network would be distributed so as to prevent highly frequent forms from invading the less frequent ones, i.e., to ensure linguistic diversity. The asymmetry evidenced by grammaticalization would thus be a consequence of the fact that the highly pervasive functional forms must be kept away from the lexical, referential, more contextspecific forms. This puzzling unidirectionality could thus have been selected as a cognitive structure able to guarantee a wide spectrum of possibilities in linguistic expression.

## 5. Sociolinguistic interpretation

We can give our model a completely different interpretation, taking a sociolinguistic view point. Instead of sites $C_{0}$ and $C_{1}$, one considers two separate communities of speakers, $C_{0}$ and $C_{1}$. Different tokens represent now different individuals, who make binary choices between either variant $X$ or variant $Y$. The different community sizes, $m$ and $M$, are then the analogous of the different memory sizes. The fact that $C_{0}$ influences unilaterally $C_{1}$ may be understood as the fact that community $C_{0}$ has some prestige compared to $C_{1}$, so that $C_{1}$ members listen to $C_{0}$ members while the reverse does not hold. Similarly, different call frequencies may represent different representations in society - people from prestige communities being given media visibility to the exclusion of
the other communities. With this purely sociolinguistic interpretation, the model formalism thus remains exactly the same. Note that this point of view is akin to the one defended in [18].

In this interpretation, however, the model does not explain why the prestige community $C_{0}$ adopted $X$ in the first place; nor does it explain the regularities in semantic change. Another point in which this interpretation weakens is the timescale. Linguistic change can be very slow, taking up to several centuries, as shown in our corpus study. Is it reasonable to presume that the social structure holds and remains the same throughout centuries? On the contrary, some aspects of conceptual structure happen to be extremely stable, as they are both deeply constitutive of a culture, e.g. through entrenched metaphors [27, and due to the generic cognitive features of the mind (expressing time relations through spatial ones [25], for instance). As it happens, metaphors prove to be very stable, even if the reasons for this stability are still unclear. The astonishing persistence of myths schemata through the ages [33] is another hint of the remarkable resilience of human cultural features.

## Appendix C: Boundaries of the trap region

The analytical definitions, used to compute the latency and growth times in the model, are based on first passage times. In this section we outline the procedure to compute these times

## 1. Analytical computation of mean first passage times

Let us note $T_{n \rightarrow m}$ the first passage time at site $m$, starting at site $n, 0 \leq n, m \leq M$. This is a random variable for which one can write down a recursion equation for its generatrix function:

$$
\begin{equation*}
\left\langle e^{\lambda T_{n \rightarrow m}}\right\rangle=R_{n}\left\langle e^{\lambda\left(T_{n+1 \rightarrow m}+1\right)}\right\rangle+L_{n}\left\langle e^{\lambda\left(T_{n-1 \rightarrow m}+1\right)}\right\rangle+\left(1-L_{n}-R_{n}\right)\left\langle e^{\lambda\left(T_{n \rightarrow m}+1\right)}\right\rangle, \tag{C1}
\end{equation*}
$$

where $R_{n}$ and $L_{n}$ are, respectively, the forward and backward jump probabilities, and $\langle$.$\rangle denotes the aver-$ age. We recall that $n=0$ is a reflecting boundary ( $L_{0}=0, R_{0}>0$ ), and $n=M$ an absorbing boundary ( $R_{M}=L_{M}=0$ ). We have $T_{n, n}=0$, and for the left boundary condition, that is for $n=0$ :
$\left\langle e^{\lambda T_{0 \rightarrow m}}\right\rangle=R_{0}\left\langle e^{\lambda\left(T_{1 \rightarrow m}+1\right)}\right\rangle+\left(1-R_{0}\right)\left\langle e^{\lambda\left(T_{0 \rightarrow m}+1\right)}\right\rangle$.
The first and second derivatives of this equation (C1) with respect to $\lambda$, at $\lambda=0$, leads to recurrence relations
for the first and second moment of $T_{n \rightarrow m}$, respectively. More specifically, we can compute the first two moments of the first passage time between one site and its immediate successor, $T_{i \rightarrow i+1}$ :

$$
\begin{equation*}
\left\langle T_{i \rightarrow i+1}\right\rangle=t_{i} \tag{C3}
\end{equation*}
$$

And:

$$
\begin{equation*}
\left\langle T_{i \rightarrow i+1}^{2}\right\rangle=u_{i} \tag{C4}
\end{equation*}
$$

Where the $t_{i}$ 's and $u_{i}$ 's are iteratively computed from:

$$
\left\{\begin{array}{l}
t_{0}=\frac{1}{R_{0}}  \tag{C5}\\
u_{0}=\frac{2 t_{0}-1}{R_{0}}
\end{array}\right.
$$

And:

$$
\left\{\begin{array}{l}
t_{i}=\frac{1}{R_{i}}+\frac{L_{i}}{R_{i}} t_{i-1}  \tag{C6}\\
u_{i}=2 t_{i}^{2}+\frac{L_{i}}{R_{i}} u_{i-1}
\end{array}\right.
$$

From this, we can easily compute the first two moments for any $T_{n \rightarrow m}$ :

$$
\begin{equation*}
\mu\left(T_{n \rightarrow m}\right)=\sum_{k=n}^{m-1} t_{k} \tag{C7}
\end{equation*}
$$

And:

$$
\begin{equation*}
\sigma^{2}\left(T_{n \rightarrow m}\right)=\sum_{k=n}^{m-1}\left(u_{k}-t_{k}^{2}\right) \tag{C8}
\end{equation*}
$$

## 2. Trap boundaries

In the main text, we explain latency time and growth time as first passage times. However, these two quantities are both empirically extracted from the macroscopic pattern obtained at the end of a run, in a procedure exactly transposed from the corpus data treatment. The question is then: Which trap boundaries $n_{\text {in }}$ and $n_{\text {out }}$ should we set in order for the properly defined time $T_{n_{\text {in }} \rightarrow n_{\text {out }}}$ correspond statistically to the empirically defined latency time?

Besides, growth time can be seen as well as a first passage time between two sites. Though the exit site should be $M$, it is more appropriate to define a cut-off $n_{\text {last }}$. Indeed, there is a discrepancy between the fact that, close to the absorbing point, the walk gets slowed down again, and that, in this region, the new variant is almost always produced anyway. In other terms, growth time, as extracted from the time evolution of the ratio of produced new variant occurrences, is not sensitive whether the end of the walk is reached or not.

Let us note $\mu_{g}$ and $\sigma_{g}^{2}$, and $\mu_{l a t}$ and $\sigma_{l a t}^{2}$, respectively the mean and the variance of the growth and latency times (obtained from the distributions of those empirically extracted quantities from ten thousand runs). Then, over a reasonable range of $n$, we look for $m$ so that $\mu\left(T_{n \rightarrow m}\right)$ is as close as possible to $\mu_{g}$; we then choose the pair $(n ; m)$ such that $\sigma^{2}\left(T_{n \rightarrow m}\right)$ is as close as possible to $\sigma_{g}^{2}$. This pair defines thus the region of growth, $\left(n_{\text {out }} ; n_{\text {last }}\right)$. We then choose $n_{\text {in }}$ so as to fit the mode of the empirical latency distribution, assuming that first passage time is distributed according to an Inverse Gaussian (and is thus a function of $\mu$ and $\sigma^{2}$ ).

## Appendix D: Growth times distribution

In the main text, Figure 7, Right panel, shows the fit of the corpus data growth times distribution by an Inverse Gaussian distribution. Here we discuss the disagreement observed for short growth times.
Because of the necessity to define the extreme values of the sigmoid, $x_{\min }$ and $x_{\max }$, a growth times of $d$ decades will be described by a number $N=d-2$ of points. The problem is, a scarce number of points is more easily compatible with any pattern than a high number of points. A linear fit of eight points is for instance more statistically significant than a linear fit of three points. For this reason, we did not consider any growth time lower than 6 decades, as stated in the main text. We kept the 6 decades growths, even though the bias is still important, as the corresponding growth time is associated with a very few number of points (4). We observe that this particular value appears over-represented compared to the Inverse Gaussian fit.

We detail below two ways by which low growth times could be over-represented in the growth times distribution. The first one is the existence of false positives. The second one accounts for the easier rejection of false negatives for higher number of points. The combination of both effects explains the spurious over-representation of six-decades growth times in our dataset. Note, however, that keeping this point allows for a better estimate of the mean growth time, which is used for the fit.

This technical issue does not affect sensitively the analysis of the simulated data, for the timescale there is such that there is no growth time associated to fewer than 8 data points.
a. False positives To give an estimate of the probability to accept false positives, we generate 100,000 samples each built through the following steps: We produce a given number $N$ of random points between 0 and 1 (the domain on which is properly scaled sigmoid takes its values), order them, apply the logit transform, and compute the $r^{2}$ parameter of the linear fit of this latter transform. We then approximate the probability of a false positive by the ratio of samples whose associated $r^{2}$ is greater than 0.98 . The probability of a false positive is significantly higher for samples of 4 points (Fig. 9). This effect is higher still if one accounts for the fact that the probability for four points to be ordered is much higher than it is for more.
b. False negatives Provided a sigmoidal pattern, what is the probability for the pattern to remains robust in presence of a Gaussian white noise? To address this question, for a given number $N$ of points, we generate a discretized sigmoid as:

$$
\begin{equation*}
y_{i}=\frac{1}{1+e^{-\alpha(N)(x-N / 2)}}, \text { for } 0 \leq i \leq N-1 \tag{D1}
\end{equation*}
$$

To provide an estimate of $\alpha(N)$, the slope of a $N$ points sigmoid, we make use of a relation exhibited by the data itself (Fig. 10), according to which $\alpha(N)=$


FIG. 9. (a) Ratio of good linear fits (i.e. such that $r^{2}>$ 0.98 ) of logit transforms of 100,000 samples of $N$ randomly generated ordered points, for different values of $N$. (b) Same ratios, divided by a factor $N$ ! to account for the ordering of the points.
$e^{3.58}(N+2 .)^{-1.47}$. Then we produce a noisy version of the sigmoid such that, for each $i, \tilde{y}_{i}=y_{i}+\eta_{i}$, with $\eta_{i}$ a white Gaussian noise of standard deviation $\sigma$. Producing 10,000 of such samples, we compute the ratio of samples such that the linear fit of their logit transforms display an $r^{2}$ greater than 0.98 . Comparing these ratios for different numbers of points lead to a bias from the uniform distribution (Fig. 11). This bias depend on the level of noise; however it is clear that when the noise is high enough, short growth times are likely to be over-represented.

## Appendix E: Further comparisons with data

In our paper, we show that an Inverse Gaussian distribution is adequate to capture both latency time and growth time distributions, indicating that these two quantities are of the same nature, and result from the same mechanism of change. However, the agreement between our model and the corpus data goes much further, as we show in this section.


FIG. 10. Log-log graph of the slope of the linear transform vs the growth time. The red line is given by the linear fit of the maximum slopes for each growth time (slope: -1.47; intercept: 3.58)


FIG. 11. Distribution over $N$ of resisting samples in presence of a Gaussian white noise of amplitude $\eta$. In absence of noise, the distribution is a uniform one.

## 1. Péclet number

The parameters $\mu$ and $\lambda$ of the Inverse Gaussian distribution scale with time length in the same way, so that is is relevant to consider their ratio, which is called the Péclet number 34. Note that, because the relation $\lambda=\mu^{3} / \sigma^{2}$ holds, the Péclet number is but the ratio between the squared mean and the variance.

The Péclet number for latency times from corpus data is equal to 1.1 while the model gives back a Péclet number of 1.4 , so they both are of the same order of magnitude. However, for growth times, we get 6.8 for corpus data, while the model gives 63 , which means that the growth time varies much less in the model than it should.

Actually, this discrepancy is rather expected. Given the definition of the Péclet number, it means that the variance of the growth time is comparatively greater in the data than it is in our model. Yet, this can be understood in terms of the latter: Indeed, it has been stressed that the conceptual network of language is organized as a small-world network 35, and we have proposed that major semantic change, characterized by the latency-growth
pattern, would correspond to a leap from a cluster to another. It means that latency involves only one bridge, so that the set-up we explored should be enough to cover it. Growth, on the other hand, depends on the cluster size, and on the inner organization of the cluster. It thus involves a varying number of contexts, which explains why the variance of the growth would be greater in actual data, leading to a smaller Péclet number.

Concerning the scale of the process, it could be tempting to compare mean latency between model and data to find the value of $M$ (size of the memory) which would correspond to the data. However, the scale entangles both $M$ and the size of the counting window. It also depends on the total number of involved contexts. There is thus no obvious way to compare the scales involved in the model and in the data.

## 2. Statistical distribution of the slopes

From the empirical procedure, we can extract, for both corpus and numerical datasets, the statistical distributions of the slopes of the logit transform of the sigmoidal part. Inverse Gaussian distribution fits very well both numerical and corpus data (Fig. 12). Gaussian distribution fits numerical data as well, but does not capture the behavior of corpus data. This is not surprising, as an Inverse Gaussian distribution tends to a Gaussian one with parameter $\lambda$ going to infinity. The fact that $\lambda$ is much bigger compared to $\mu$ in numerical data than in corpus data is in agreement with the discrepancy between the Péclet numbers for the growth time distributions: There are more sources of variation for the growth part of the process than what we considered in the model. However, we still have no definitive explanation to provide concerning the fact that slopes should follow an Inverse Gaussian distribution as well.

## 3. Latency-Growth correlation

It may be intuitively expected for latency and growth to be correlated: The longer the wait, the more momentum is gained. Yet, according to our model, there is no such correlation: Latency and growth times, as seen as first passage times in different parts of a Markov chain, are strictly independent quantities. However, in the empirical procedure, these two parameters become correlated, for the latency is defined as the time spent in a region comprised between $x_{t_{\text {out }}} \pm a *\left(1-x_{t_{o u t}}\right)$, where $x_{t_{\text {out }}}$ is the frequency attained at the beginning of the growth process and $a$ is set to 0.17 . Thus, the higher $x_{t_{\text {out }}}$, the smaller the margin, so that a short growth (high $x_{t_{\text {out }}}$ ) will be correlated with a short latency. These two quantities are thus weakly positively correlated, with a Pearson coefficient of 0.20 (Fig. 13b).

If we now turn to corpus data, we find a Pearson coefficient of 0.24 (Fig 13a). The correlation between latency


FIG. 12. Statistical distribution of the slopes of the logit transform of the sigmoidal growth, for both corpus data (a) and numerical data (b). An Inverse Gaussian distribution fits both.
and growth is weak, and can be entirely imputed to the details of the empirical procedure, as we have just seen for the numerical data. It thus means that growth time and latency time are two independent quantities, so that positing a Markovian nature of language change is in line with findings from corpus data.

## 4. Growth-Slope correlation

Growth and slope are expected to be correlated. One may even expect to find a scaling law between the two, in line with what has been evidenced for other socio-cultural changes [36]. We did not find any striking scaling law, yet the two quantities are convincingly negatively correlated, both in our model (Pearson coefficient of -0.69 , Fig. 14b) and in corpus data (Pearson coefficient of -0.57 , Fig 14a).

(a)

(b)

FIG. 13. Pearson coefficient for the correlation between growth time and latency time obtained from (a) corpus data and (b) numerical simulations

## Appendix F: Why not using Google Ngram?

Google Ngram (https:/books.google.com/ngrams) gathers an impressive quantity of digitalized books from about the sixteenth century. It hosts about 800,000 texts in French (about twenty times more than Frantext). Nevertheless, it presents some major limitations which make this database inappropriate for the present study, as we discuss at length in this section.

Some limits of the Google Ngram database have already been stressed in the recent past [37]. However, these concerns are specifically relevant for lexical changes, most subject to socio-historical contingencies, and they do not straightforwardly apply to our aims. Functional words, unlike proper names like 'Frodo' or items like 'computer', are not that sensitive to cultural shifts. This is why we point out other serious limits inherent to Google Ngram. In the following, we use Google Books as a probe to the contents of Google Ngram,

(a)

(b)

FIG. 14. Pearson coefficient for the correlation between growth time and slope obtained from (a) corpus data and (b) numerical simulations
though the two algorithms are different (e.g. the former does not regognize ponctuation while the latter does), and the exact overlapping between the contents of the two databases is unknown.

The first concern about Google Ngram regards quality of digitalization. Texts older than the nineteenth century have been printed in fonts for which the character recognition algorithm has clearly not been optimized. For instance, the following sentence from The royal dictionary abridged, in two parts, by Abel Boyer, 1715: 'Parler avantageusement de quelqu'un, to speak well of one, to speak much to his advantage, to give a good character of him, to speak honourably of him.' has been transcribed as: 'Parler avantageusement e quelqu'un, $1^{\circ}$ speak well of one, te steak much to his advant 1ge, to ive a gead characier of him, to steak h2nourably of him.' Some words, such as 'steak' and 'rince', consequently appear much more frequently than they should, as they are mistaken for 'speak' and 'Prince'. Another example of this poor scan-
ning quality can be seen in the comparison between: 'I found that the New-modelling of this Story, would force me sometimes on the difficult Task of making the chiefest Persons speak something like their Characters, on Matter whereof I had no Ground in my Author.' and 'I faura that the Ne: -we kling of this Story, troi'i fr e ve \{ ctives on the di ili 7 k of making ti e li fist Perffns steak \{ like their Carefiers, en -i/attro sviereof. I had no Gréard in , Author.', to be found in The History of King Lear, A Tragedy. Acted as the King's-Theatre. by Nahum Tate, 1736. The original text is admittedly hard to decipher, yet any posterior check on the scan would immediately detect such nonsensical concatenations of characters. By comparison, every text in the Frantext database has been digitalized with great care and such blatant errors are not to be found.

The second point on which Frantext deserves to be favored is the kind of available data. Google Ngram provides statistics on n-grams, which are strings of n successive items (the so-called 'grams'), with n ranging from 1 to 5 . For each n-gram, it is provided, per year, the number of times it appears and in how many texts. Thus one cannot identify in which texts it appears most; nor can one have access to its context of use. The only way to probe the contents of Google Ngram is through Google Books (which we used here for all the discussed examples), yet it seems impossible to know the exact overlap between the two databases. This data structure based on n-grams is furthermore limiting when it comes to slot constructions. For instance, the French construction 'à X reprises', with X being a quantity, can hardly be tracked using Google Ngram, as it corresponds to far too many ngrams, which would need to be listed one by one: 'à deux reprises', ‘à deux ou trois reprises', ‘à plusieurs reprises', 'à de nombreuses reprises', etc. This search is made all the more difficult by the fact that 'à' did not always take an accent in older texts. In Frantext, as we have seen in Appendix A we can work out an elaborate request using booleans and blank words to capture the diverse uses of this construction and overcome the orthographic difficulties.

The third and final point we want to stress here is the choice of texts and their dating. In Frantext, a text may appear in several editions, as is the case for Le Cid, by Pierre Corneille, which appear thrice in the database, associated to the years 1637, 1637 and 1682. These dates usually correspond to the first edition of a book, rather than to the edition which is actually digitalized (such information being also provided). Google Ngram displays about thirty versions of Le Cid, with publication ranging from 1775 to 2013, some of them being ascribed to Jean Racine (as they are found in several editions of a book called Oeuvres de J. Racine et de P. et T. Corneille). The case of Le Cid is, in Frantext, quite an exception; while in Google Ngram, most famous classical novels from past centuries are found in a dozen versions at least.

The contents of the database is problematic as well. As highlighted in 37, Google Ngram over-represent aca-
demic literature. This also tends to bias the data; for instance, among the fourteen results of the request 'par ma barbe' on the French Google Books subdatabase, for the years 1950-2000, only three of them are relevant, two being modern translations of older texts (Don Quixote and an nineteenth century German play by Töpffer). The third one comes from an anthology of French folktales. All other occurrences are academic quotes and glosses of past works, or reprints of such works. In such a case, it means than only one fifth of the occurrences would be reliable as a reflect of language use in this time period (two of them being borderline cases). Frantext, on the other hand, has two occurrences of 'par ma barbe', one of them from the song lyrics of singer Georges Brassens, the other from a 1988 translation of a Shakespeare play (and so more debatable). There is thus almost as many relevant occurrences in Frantext and Google Ngram (two versus three), while none in Frantext are completely irrelevant.

This being said, Google Ngram is a formidable tool, which can lead to interesting insights and be of great use. It is not, however, fitted for the work that we performed, where we need an accuracy and a reliability that this database is unable to provide.

## Appendix G: Studied forms

The corpus data through which we evidence the latency-growth pattern has been extracted from the study of about 400 hundred semantic expansions in the functional realm (with the exception of liberté, which we have shown to suggest the further generality of the pattern), with the help of the Frantext database and retrieving tools. We selected these forms according to several criteria: They must have undergone at least one semantic expansion towards a functional use during the time period under consideration; they must be easily distinguished (e.g. entre deux, in the meaning of 'in between', can be confused with occurrences of literal meaning 'between two'). The set of chosen forms is far from exhausting the pool of possible examples.

On table we provide the full list of studied forms. For each of those, we display the length (in decades) of the latency part, of the growth part, the slope of the logit transform of the growth part, the $r^{2}$ parameter associated to the linear fit of this logit transform, the standard deviation between the data and the inferred sigmoidal pattern, and the total number of occurrences of the form in our corpus. A standard deviation less than 0.03 corresponds to an especially good sigmoidal fit of the data. Some forms are listed several times; it corresponds to the case where a form underwent several semantic expansion processes, each associated with the latency-growth pattern. 'BUG' corresponds to a flaw of the Frantext search and retrieve algorithm, sometimes unable to build up the output file of the query. This bug cannot be override through a manual manoeuvre, for it is caused by a faulty
encoding of some parts of the texts. The data thus exist, but could not be retrieved.

An upper-case 'NO' indicates that no such pattern has been found in the time-evolution of the frequency of that form. The fact that a form does not follow a S-curve during its semantic expansion may spread doubt on the genericity of this pattern. In many cases however, we found such a pattern and rejected it because the data was too spurious. It should also be stressed that the criterion we applied (the $r^{2}$ residual of the linear fit of the logit transform must be greater than 0.98 , see Materials and Methos) is a really strong one and can easily lead to rejection.

It is nonetheless interesting to note that the robustness of the pattern does not depend excessively on the scarcity of data. Indeed, instances associated to a very low number of occurrences can lead to a very clean pattern (e.g. à plus d'un titre, whose growth lasts for 8 decades in total, scores as low as 59 occurrences, and still brings out
a remarkable $r^{2}$ of 0.995). What seems to be crucial is thus not the question of how much data we can get, but of whether or not the change is isolated. Indeed, some changes are not independent from one another. Many constructions beginning with the preposition par, for instance, follow their own course of evolution, while the meaning of par itself also expands. Several constructions can also compete for the same paradigm (e.g. il me semble, je pense, je suppose). Their individual frequency pattern not following an S-curve of growth may thus be seen as resulting from interferences between the different semantic expansion processes. In these cases, only the refinement of linguistic queries can lead to better results. It thus confirms, once again, the necessity to rely on a clean and easily manipulable database rather than on giant databases where the sheer amount of data is of no help.

TABLE I: List of studied forms

| Researched form | Latency | Growth | Slope | $\mathbf{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| à base de | 9 | 8 | 1.02 | 0.983 | 0.043 | 607 |
| à bien des égards (i) | 0 | 8 | 0.79 | 0.983 | 0.049 | 147 |
| à bien des égards (ii) | 2 | 6 | 1.57 | 0.984 | 0.041 | 147 |
| à bord de | NON | NON | NON | NON | NON | 1728 |
| acabit | 0 | 7 | 1.20 | 0.992 | 0.031 | 148 |
| à cause de | NON | NON | NON | NON | NON | 24840 |
| à cause que | 2 | 6 | 0.61 | 0.997 | 0.105 | 2516 |
| à ce moment | 14 | 13 | 0.46 | 0.990 | 0.080 | 8861 |
| à ce propos (i) | 2 | 7 | 2.22 | 0.988 | 0.011 | 1711 |
| à ce propos (ii) | 13 | 7 | 0.81 | 0.991 | 0.048 | 1711 |
| à ce sujet | 10 | 7 | 1.95 | 0.988 | 0.037 | 4001 |
| à cet égard | 0 | 6 | 0.85 | 0.993 | 0.087 | 4974 |
| à cet instant | 2 | 13 | 0.55 | 0.993 | 0.034 | 1198 |
| à condition de | 5 | 9 | 0.79 | 0.991 | 0.035 | 1151 |
| à condition que (i) | 11 | 6 | 1.19 | 0.997 | 0.031 | 1653 |
| à condition que (ii) | 4 | 6 | 0.83 | 0.981 | 0.068 | 1653 |
| à contre-courant | 3 | 11 | 0.75 | 0.995 | 0.021 | 171 |
| à côté de | 23 | 13 | 0.52 | 0.990 | 0.034 | 18065 |
| à coup sûr | 7 | 7 | 1.70 | 0.996 | 0.012 | 2546 |
| à court terme | 13 | 7 | 2.19 | 0.997 | 0.019 | 751 |
| à couvert | NON | NON | NON | NON | NON | 1144 |
| actuellement | 10 | 11 | 0.46 | 0.989 | 0.048 | 6618 |
| à découvert | 1 | 7 | 1.33 | 0.981 | 0.035 | 930 |
| à défaut de | NON | NON | NON | NON | NON | 1725 |
| afin de | 4 | 6 | 0.81 | 0.995 | 0.073 | 21833 |
| afin que | BUG | BUG | BUG | BUG | BUG | 19850 |
| à fond de | 8 | 6 | 0.91 | 0.987 | 0.059 | 486 |
| à fond de train | BUG | BUG | BUG | BUG | BUG | 180 |
| à force | NON | NON | NON | NON | NON | 294 |
| à force de | NON | NON | NON | NON | NON | 8178 |
| à grand renfort de | NON | NON | NON | NON | NON | 230 |
| ainsi donc | NON | NON | NON | NON | NON | 1247 |
| à la base | NON | NON | NON | NON | NON | 574 |
| à l'accoutumée | 0 | 8 | 0.99 | 0.988 | 0.027 | 196 |
| à l'aide de | NON | NON | NON | NON | NON | 5247 |
| à la limite | 7 | 11 | 0.73 | 0.983 | 0.076 | 603 |
| à la lisière de | 6 | 12 | 0.56 | 0.985 | 0.036 | 527 |
| à la longue | 9 | 7 | 0.57 | 0.987 | 0.115 | 1245 |


| Researched form | Latency | Growth | Slope | $\mathbf{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| à la lumière de | 0 | 7 | 1.10 | 0.987 | 0.031 | 1141 |
| à la mesure de | 24 | 9 | 1.14 | 0.988 | 0.056 | 820 |
| à la place | 22 | 22 | 0.31 | 0.983 | 0.029 | 5638 |
| à la rigueur | 9 | 8 | 1.14 | 0.983 | 0.049 | 1717 |
| à l'écart | NON | NON | NON | NON | NON | 2517 |
| à l'écart de | 6 | 12 | 0.58 | 0.992 | 0.030 | 854 |
| à l'égard de | 7 | 11 | 1.19 | 0.992 | 0.038 | 13396 |
| à l'encontre de | 9 | 17 | 0.45 | 0.986 | 0.026 | 1272 |
| à l'envi | 0 | 9 | 0.88 | 0.991 | 0.025 | 817 |
| à l'exception de | 1 | 8 | 0.81 | 0.995 | 0.080 | 1883 |
| à l'heure actuelle | 0 | 11 | 0.95 | 0.981 | 0.045 | 858 |
| à l'heure dite | NON | NON | NON | NON | NON | 234 |
| à l'heure où | 2 | 9 | 0.79 | 0.982 | 0.043 | 1779 |
| à l'improviste | NON | NON | NON | NON | NON | 1024 |
| à l'instant | 0 | 6 | 1.02 | 0.993 | 0.047 | 1550 |
| à l'instar de (i) | 4 | 10 | 0.61 | 0.981 | 0.044 | 663 |
| à l'instar de (ii) | 1 | 7 | 0.99 | 0.988 | 0.035 | 663 |
| à l'insu | 0 | 22 | 0.36 | 0.982 | 0.045 | 2776 |
| à l'inverse | 8 | 10 | 1.06 | 0.988 | 0.021 | 764 |
| à l'occasion de | 6 | 8 | 1.52 | 0.983 | 0.042 | 2032 |
| à l'orée de | 5 | 6 | 1.00 | 0.996 | 0.047 | 311 |
| alors que (i) | 3 | 7 | 1.01 | 0.983 | 0.044 | 28016 |
| alors que (ii) | 4 | 13 | 0.50 | 0.983 | 0.031 | 28016 |
| à mesure de | 4 | 8 | 0.71 | 0.990 | 0.087 | 774 |
| à mesure que | 12 | 7 | 1.74 | 0.993 | 0.017 | 10183 |
| à moins que | 2 | 8 | 0.90 | 0.981 | 0.059 | 5924 |
| à mon avis | NON | NON | NON | NON | NON | 1989 |
| à nouveau | 7 | 13 | 0.59 | 0.987 | 0.031 | 6039 |
| à outrance | 1 | 6 | 0.53 | 0.996 | 0.132 | 552 |
| à part | 0 | 28 | 0.27 | 0.986 | 0.037 | 12506 |
| à part entière | 0 | 8 | 1.33 | 0.983 | 0.034 | 180 |
| à partir de | 0 | 9 | 0.76 | 0.986 | 0.038 | 10996 |
| à peine (i) | 0 | 6 | 1.73 | 0.994 | 0.023 | 40230 |
| à peine (ii) | 0 | 6 | 1.70 | 0.986 | 0.026 | 40230 |
| à peu de chose près | 0 | 7 | 0.94 | 0.987 | 0.039 | 320 |
| à plus d'un titre | 1 | 7 | 1.02 | 0.995 | 0.031 | 59 |
| à plusieurs reprises | 9 | 7 | 1.22 | 0.994 | 0.023 | 3873 |
| après ce | NON | NON | NON | NON | NON | 101 |
| après que | 6 | 6 | 2.34 | 0.997 | 0.022 | 8487 |
| après quoi | 10 | 16 | 0.53 | 0.982 | 0.048 | 3468 |
| après tout | NON | NON | NON | NON | NON | 7741 |
| a priori | 3 | 9 | 1.21 | 0.985 | 0.046 | 1565 |
| à propos | NON | NON | NON | NON | NON | 1255 |
| à propos de | 0 | 12 | 0.50 | 0.980 | 0.056 | 9414 |
| à proprement parler | NON | NON | NON | NON | NON | 1204 |
| à rebours (i) | 2 | 6 | 1.09 | 0.994 | 0.070 | 640 |
| à rebours (ii) | 2 | 6 | 0.85 | 0.997 | 0.068 | 640 |
| à qui mieux mieux | NON | NON | NON | NON | NON | 247 |
| à sa guise | 0 | 6 | 1.40 | 0.992 | 0.028 | 1079 |
| à son terme | 1 | 11 | 0.75 | 0.991 | 0.032 | 359 |
| à tel point que (i) | 0 | 7 | 0.75 | 0.996 | 0.054 | 555 |
| à tel point que (ii) | 0 | 6 | 1.69 | 0.991 | 0.028 | 555 |
| à terme | 12 | 6 | 0.56 | 0.997 | 0.117 | 470 |
| à titre de | 5 | 13 | 0.63 | 0.990 | 0.026 | 1481 |
| à tous égards | 0 | 6 | 1.91 | 0.998 | 0.013 | 556 |
| à tout à l'heure | 0 | 10 | 0.93 | 0.983 | 0.037 | 280 |
| à tout instant | NON | NON | NON | NON | NON | 903 |
| à tout moment | 5 | 6 | 1.49 | 0.988 | 0.032 | 2262 |
| à tout prendre | NON | NON | NON | NON | NON | 480 |


| Researched form | Latency | Growth | Slope | $\mathrm{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| au bord de | NON | NON | NON | NON | NON | 11850 |
| au bout de | NON | NON | NON | NON | NON | 23173 |
| au bout du compte | NON | NON | NON | NON | NON | 469 |
| au contraire | 5 | 8 | 1.13 | 0.990 | 0.027 | 29571 |
| au contraire de | 1 | 8 | 1.14 | 0.989 | 0.028 | 1429 |
| aucunefois | BUG | BUG | BUG | BUG | BUG | 1248 |
| au demeurant | 0 | 12 | 0.68 | 0.983 | 0.039 | 1344 |
| au dépourvu | NON | NON | NON | NON | NON | 402 |
| au détriment de | 5 | 6 | 1.03 | 0.981 | 0.050 | 798 |
| au dernier moment | NON | NON | NON | NON | NON | 1370 |
| au final | NON | NON | NON | NON | NON | 38 |
| au fur et à mesure | 6 | 12 | 0.72 | 0.987 | 0.027 | 1908 |
| au jour d'aujourd'hui | NON | NON | NON | NON | NON | 87 |
| au même moment | 2 | 6 | 0.75 | 0.992 | 0.081 | 1437 |
| au moment où | 6 | 19 | 0.49 | 0.984 | 0.024 | 12729 |
| à un moment donné | 1 | 12 | 0.48 | 0.980 | 0.043 | 659 |
| au passage | 0 | 7 | 1.43 | 0.990 | 0.034 | 1754 |
| au pire | 0 | 6 | 1.63 | 0.994 | 0.026 | 401 |
| au reste | 0 | 7 | 1.39 | 0.987 | 0.036 | 4375 |
| au sujet de | 4 | 9 | 0.87 | 0.986 | 0.035 | 4945 |
| au terme de | 3 | 6 | 0.66 | 0.991 | 0.102 | 1492 |
| aux trousses | NON | NON | NON | NON | NON | 419 |
| avant tout | 27 | 10 | 0.91 | 0.986 | 0.029 | 5342 |
| avec force | NON | NON | NON | NON | NON | 324 |
| bah | 9 | 9 | 1.31 | 0.992 | 0.021 | 2681 |
| bien entendu | 33 | 18 | 0.44 | 0.984 | 0.042 | 4476 |
| bien sûr | 5 | 9 | 0.74 | 0.994 | 0.029 | 7997 |
| bref | 12 | 7 | 1.07 | 0.993 | 0.033 | 5536 |
| brusquement | 11 | 8 | 1.69 | 0.993 | 0.029 | 1783 |
| carrément (i) | 1 | 8 | 1.22 | 0.982 | 0.044 | 1207 |
| carrément (ii) | 1 | 7 | 1.55 | 0.982 | 0.042 | 1207 |
| ce faisant (i) | 0 | 6 | 1.88 | 0.992 | 0.027 | 781 |
| ce faisant (ii) | 19 | 8 | 0.64 | 0.994 | 0.059 | 781 |
| ce par quoi | 0 | 10 | 0.61 | 0.984 | 0.040 | 163 |
| c'est alors que | 6 | 10 | 0.85 | 0.982 | 0.026 | 3223 |
| c'est pour le coup que | 0 | 6 | 0.77 | 0.990 | 0.079 | 64 |
| c'est pourquoi (i) | 0 | 13 | 0.56 | 0.984 | 0.053 | 10994 |
| c'est pourquoi (ii) | 0 | 8 | 0.95 | 0.986 | 0.040 | 10994 |
| chemin faisant | NON | NON | NON | NON | NON | 641 |
| complètement | NON | NON | NON | NON | NON | 11560 |
| compte tenu de | 0 | 8 | 1.26 | 0.985 | 0.038 | 928 |
| concernant | 9 | 10 | 1.10 | 0.984 | 0.047 | 3477 |
| considérant que | NON | NON | NON | NON | NON | 191 |
| contre mon attente | NON | NON | NON | NON | NON | 102 |
| contre toute attente | 0 | 6 | 0.84 | 0.991 | 0.080 | 167 |
| d'abord et avant tout (i) | 0 | 6 | 0.80 | 0.982 | 0.073 | 62 |
| d'abord et avant tout (ii) | 1 | 6 | 1.24 | 0.982 | 0.039 | 62 |
| dans ce cas | NON | NON | NON | NON | NON | 4289 |
| dans la mesure de | NON | NON | NON | NON | NON | 480 |
| dans la mesure du possible | 0 | 11 | 0.62 | 0.988 | 0.050 | 188 |
| dans la mesure où | NON | NON | NON | NON | NON | 2753 |
| dans le cadre de | 11 | 6 | 1.27 | 0.988 | 0.069 | 1145 |
| dans le même temps (i) | 0 | 9 | 1.02 | 0.986 | 0.029 | 1217 |
| dans le même temps (ii) | 3 | 7 | 0.90 | 0.983 | 0.046 | 1217 |
| dans l'ensemble (i) | 0 | 8 | 0.99 | 0.986 | 0.045 | 1809 |
| dans l'ensemble (ii) | 10 | 7 | 0.82 | 0.981 | 0.068 | 1809 |
| dans l'immédiat | 10 | 9 | 1.10 | 0.984 | 0.033 | 329 |
| dans quelque temps (i) | 0 | 6 | 0.81 | 0.987 | 0.081 | 234 |
| dans quelque temps (ii) | 0 | 6 | 0.76 | 0.991 | 0.077 | 234 |


| Researched form | Latency | Growth | Slope | $\mathbf{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| dans son ensemble | 1 | 8 | 0.67 | 0.990 | 0.063 | 835 |
| dans un autre temps | NON | NON | NON | NON | NON | 143 |
| dans un cas comme dans l'autre | 1 | 8 | 1.04 | 0.983 | 0.042 | 111 |
| dans une large mesure | 5 | 8 | 0.66 | 0.994 | 0.062 | 381 |
| dans un instant | 2 | 10 | 0.87 | 0.981 | 0.035 | 661 |
| dans un moment | 0 | 15 | 0.47 | 0.984 | 0.037 | 1473 |
| dans un premier temps | NON | NON | NON | NON | NON | 229 |
| dans tous les cas | NON | NON | NON | NON | NON | 1609 |
| d'autant plus | 0 | 9 | 0.52 | 0.987 | 0.061 | 11584 |
| d'autant plus que | NON | NON | NON | NON | NON | 3339 |
| d'autre part | 24 | 12 | 0.64 | 0.982 | 0.045 | 11012 |
| décidément | 2 | 13 | 0.50 | 0.984 | 0.048 | 4795 |
| de ce fait | 2 | 8 | 0.66 | 0.993 | 0.105 | 628 |
| de façon que | NON | NON | NON | NON | NON | 1473 |
| de fait | 0 | 8 | 0.91 | 0.989 | 0.034 | 5018 |
| d'année en année | NON | NON | NON | NON | NON | 505 |
| de ce côté | NON | NON | NON | NON | NON | 3665 |
| de jour en jour | NON | NON | NON | NON | NON | 2217 |
| de la part de | NON | NON | NON | NON | NON | 16400 |
| de la sorte | 11 | 8 | 0.77 | 0.982 | 0.046 | 3752 |
| de l'aveu de | 0 | 17 | 0.34 | 0.986 | 0.054 | 196 |
| de l'avis de | 0 | 6 | 1.17 | 0.987 | 0.039 | 146 |
| de loin | 16 | 10 | 0.73 | 0.994 | 0.031 | 1262 |
| de loin en loin | 0 | 17 | 0.41 | 0.984 | 0.042 | 1348 |
| de long en large | 3 | 9 | 0.76 | 0.983 | 0.054 | 734 |
| de main en main | NON | NON | NON | NON | NON | 464 |
| d'emblée | 3 | 10 | 0.72 | 0.986 | 0.034 | 1451 |
| de mèche | 0 | 6 | 0.72 | 0.987 | 0.087 | 98 |
| de mieux en mieux | 0 | 6 | 1.32 | 0.996 | 0.028 | 445 |
| de moins en moins | 6 | 21 | 0.28 | 0.980 | 0.038 | 1536 |
| de mon côté | 0 | 14 | 0.71 | 0.981 | 0.043 | 8788 |
| de mon fait | NON | NON | NON | NON | NON | 467 |
| de nulle part | 24 | 12 | 0.36 | 0.993 | 0.063 | 289 |
| de pair | 12 | 7 | 1.04 | 0.983 | 0.050 | 578 |
| de place en place | 13 | 7 | 0.86 | 0.988 | 0.058 | 376 |
| de point en point | 0 | 6 | 0.75 | 0.988 | 0.079 | 247 |
| de part en part (i) | 0 | 7 | 1.49 | 0.995 | 0.019 | 498 |
| de part en part (ii) | 7 | 6 | 1.42 | 0.989 | 0.027 | 498 |
| de part et d'autre | NON | NON | NON | NON | NON | 2505 |
| de plus en plus | 0 | 6 | 1.08 | 0.999 | 0.038 | 18226 |
| de près ou de loin | NON | NON | NON | NON | NON | 221 |
| de proche en proche | 3 | 9 | 1.00 | 0.987 | 0.024 | 702 |
| de quelque part | NON | NON | NON | NON | NON | 166 |
| des fois | 5 | 10 | 0.89 | 0.982 | 0.032 | 1423 |
| des fois que | 2 | 6 | 0.90 | 0.988 | 0.063 | 182 |
| dès l'instant | 1 | 8 | 0.96 | 0.988 | 0.030 | 769 |
| dès lors que | 21 | 7 | 0.59 | 0.983 | 0.126 | 994 |
| de sorte que | 6 | 6 | 0.85 | 0.997 | 0.063 | 11320 |
| de surcrot | 38 | 9 | 0.91 | 0.988 | 0.032 | 720 |
| de temps à autre | 17 | 14 | 0.74 | 0.983 | 0.028 | 3547 |
| de temps en temps | 2 | 13 | 0.45 | 0.981 | 0.047 | 8916 |
| de toute façon | 35 | 7 | 0.94 | 0.996 | 0.066 | 3595 |
| de toute manière | NON | NON | NON | NON | NON | 727 |
| de toutes façons (i) | 4 | 6 | 2.04 | 0.989 | 0.039 | 715 |
| de toutes façons (ii) | 1 | 6 | 1.56 | 0.986 | 0.038 | 715 |
| de toutes parts | 0 | 8 | 1.38 | 0.992 | 0.035 | 4792 |
| d'heure en heure | NON | NON | NON | NON | NON | 573 |
| d'ici là | 16 | 9 | 0.60 | 0.985 | 0.096 | 904 |
| dorénavant | NON | NON | NON | NON | NON | 256 |


| Researched form | Latency | Growth | Slope | $\mathrm{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| d'outre en outre | NON | NON | NON | NON | NON | 47 |
| du fait de | 24 | 8 | 0.73 | 0.986 | 0.055 | 1423 |
| du même coup | 14 | 17 | 0.47 | 0.991 | 0.026 | 1502 |
| du moment que | 3 | 6 | 1.64 | 0.999 | 0.015 | 1765 |
| d'une manière ou d'une autre | NON | NON | NON | NON | NON | 320 |
| d'une part (i) | 0 | 8 | 1.13 | 0.985 | 0.038 | 5671 |
| d'une part (ii) | 3 | 7 | 0.80 | 0.982 | 0.061 | 5671 |
| d'une voix claire | 6 | 10 | 0.68 | 0.993 | 0.030 | 13511 |
| du pareil au même | NON | NON | NON | NON | NON | 92 |
| du point de vue de | 7 | 8 | 0.80 | 0.995 | 0.039 | 899 |
| du reste | 1 | 6 | 1.82 | 0.994 | 0.027 | 5510 |
| en attendant | NON | NON | NON | NON | NON | 3351 |
| en attendant de | 0 | 6 | 1.41 | 0.983 | 0.038 | 510 |
| en attendant que | NON | NON | NON | NON | NON | 2270 |
| en bordure de | 0 | 11 | 0.83 | 0.983 | 0.030 | 434 |
| en bref | 1 | 6 | 1.06 | 0.996 | 0.041 | 339 |
| en ce moment (i) | 5 | 8 | 1.09 | 0.985 | 0.038 | 12751 |
| en ce moment (ii) | 2 | 9 | 0.79 | 0.983 | 0.038 | 12751 |
| en ce que | 0 | 7 | 1.55 | 0.990 | 0.036 | 3971 |
| en ce qui concerne | 34 | 8 | 0.70 | 0.981 | 0.057 | 3950 |
| en ce qui me concerne | NON | NON | NON | NON | NON | 682 |
| en considération de | NON | NON | NON | NON | NON | 409 |
| en cours de | 0 | 14 | 0.54 | 0.984 | 0.074 | 1110 |
| en cours de route | 0 | 8 | 0.91 | 0.986 | 0.036 | 301 |
| en d'autres termes (i) | 0 | 6 | 0.63 | 0.980 | 0.102 | 1228 |
| en d'autres termes (ii) | 0 | 9 | 0.66 | 0.987 | 0.058 | 1228 |
| en définitive | NON | NON | NON | NON | NON | 1538 |
| en dépit de | NON | NON | NON | NON | NON | 4016 |
| en face de | 18 | 17 | 0.39 | 0.982 | 0.058 | 10956 |
| en façon que | NON | NON | NON | NON | NON | 48 |
| en fait | 19 | 11 | 0.57 | 0.985 | 0.047 | 8871 |
| en fin de compte | 27 | 13 | 0.42 | 0.980 | 0.054 | 1417 |
| en gros | 18 | 7 | 1.06 | 0.984 | 0.044 | 320 |
| en guise de | 7 | 10 | 0.93 | 0.982 | 0.035 | 1598 |
| en instance de | NON | NON | NON | NON | NON | 77 |
| en l'occurrence | 0 | 11 | 0.58 | 0.993 | 0.038 | 525 |
| en long et en large | NON | NON | NON | NON | NON | 108 |
| en même temps | 3 | 10 | 0.88 | 0.996 | 0.023 | 18370 |
| en même temps que | NON | NON | NON | NON | NON | 8241 |
| en mesure de | NON | NON | NON | NON | NON | 1470 |
| en particulier (i) | 2 | 6 | 1.26 | 0.993 | 0.050 | 8949 |
| en particulier (ii) | 17 | 15 | 0.39 | 0.984 | 0.057 | 8949 |
| en particulier (iii) | 4 | 7 | 0.80 | 0.987 | 0.072 | 8949 |
| en partie | NON | NON | NON | NON | NON | 5645 |
| en passe de | NON | NON | NON | NON | NON | 46 |
| en plein | 7 | 6 | 0.51 | 0.997 | 0.135 | 183 |
| en plein qqch | 31 | 10 | 0.86 | 0.982 | 0.035 | 15939 |
| en quelque sorte | NON | NON | NON | NON | NON | 6422 |
| en sorte que | 3 | 6 | 0.97 | 0.982 | 0.056 | 4786 |
| en suspens | BUG | BUG | BUG | BUG | BUG | 961 |
| en tant que tel | 8 | 6 | 0.91 | 0.997 | 0.065 | 314 |
| entre autres | NON | NON | NON | NON | NON | 4402 |
| en vérité (i) | 0 | 6 | 0.93 | 0.988 | 0.060 | 8194 |
| en vérité (ii) | 1 | 6 | 0.75 | 0.981 | 0.080 | 8194 |
| en voie de (i) | 2 | 6 | 0.57 | 0.996 | 0.200 | 1027 |
| en voie de (ii) | 21 | 22 | 0.31 | 0.980 | 0.052 | 1027 |
| en vue de | 5 | 6 | 1.31 | 0.995 | 0.026 | 3625 |
| époque | 7 | 14 | 0.75 | 0.993 | 0.037 | 32290 |
| essentiellement | NON | NON | NON | NON | NON | 5471 |


| Researched form | Latency | Growth | Slope | $\mathrm{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| étant donné que | 2 | 13 | 0.62 | 0.984 | 0.036 | 341 |
| et après | NON | NON | NON | NON | NON | 7562 |
| excepté | 5 | 7 | 0.66 | 0.990 | 0.076 | 5042 |
| faute de (i) | 5 | 7 | 1.76 | 0.990 | 0.024 | 6725 |
| faute de (ii) | 4 | 11 | 0.58 | 0.983 | 0.054 | 6725 |
| faute de quoi | NON | NON | NON | NON | NON | 262 |
| force est de | NON | NON | NON | NON | NON | 84 |
| fors | BUG | BUG | BUG | BUG | BUG | 4451 |
| graduellement | 3 | 9 | 1.42 | 0.983 | 0.061 | 827 |
| hormis | 5 | 6 | 1.01 | 0.987 | 0.070 | 1464 |
| il me semble | NON | NON | NON | NON | NON | 1822 |
| il s'agit de | 3 | 11 | 0.50 | 0.983 | 0.044 | 11558 |
| il y a moyen | 4 | 6 | 1.34 | 0.994 | 0.088 | 1295 |
| j'ai l'impression | 0 | 9 | 0.57 | 0.983 | 0.052 | 74 |
| ja soit ce que | NON | NON | NON | NON | NON | 268 |
| je pense | 5 | 6 | 1.48 | 0.989 | 0.029 | 4033 |
| je suppose | 0 | 8 | 1.00 | 0.995 | 0.025 | 1110 |
| j'imagine | 11 | 9 | 0.39 | 0.989 | 0.107 | 824 |
| jusque là | 0 | 6 | 0.81 | 0.980 | 0.075 | 6908 |
| juste un | 14 | 8 | 0.83 | 0.984 | 0.063 | 1366 |
| l'autre jour | NON | NON | NON | NON | NON | 4438 |
| lendemain | NON | NON | NON | NON | NON | 28780 |
| le temps de | 20 | 11 | 0.54 | 0.987 | 0.035 | 1195 |
| liberté | 2 | 9 | 0.87 | 0.990 | 0.031 | 46705 |
| l'un dans l'autre | NON | NON | NON | NON | NON | 69 |
| l'un après l'autre | NON | NON | NON | NON | NON | 2010 |
| m'est avis | NON | NON | NON | NON | NON | 797 |
| nettement | 0 | 7 | 0.68 | 0.982 | 0.070 | 6109 |
| nommément | 1 | 6 | 1.20 | 0.981 | 0.044 | 453 |
| non pas tant | 3 | 6 | 1.25 | 1.000 | 0.029 | 855 |
| non seulement | NON | NON | NON | NON | NON | 22599 |
| non pas seulement | 6 | 6 | 0.88 | 0.988 | 0.062 | 1605 |
| notamment | 10 | 8 | 0.55 | 0.991 | 0.091 | 7508 |
| nulle part | 5 | 12 | 0.57 | 0.980 | 0.045 | 5006 |
| or donc | 4 | 6 | 2.50 | 0.988 | 0.028 | 237 |
| ouille | 1 | 7 | 1.36 | 0.997 | 0.015 | 106 |
| outre mesure | 3 | 7 | 0.86 | 0.997 | 0.054 | 664 |
| par à-coups | 0 | 11 | 0.55 | 0.985 | 0.042 | 212 |
| par ailleurs | 27 | 11 | 0.93 | 0.983 | 0.035 | 2676 |
| par avance | 2 | 12 | 0.64 | 0.980 | 0.048 | 1265 |
| par ce fait | 0 | 9 | 1.05 | 0.990 | 0.028 | 101 |
| par conséquent | 6 | 6 | 1.30 | 0.998 | 0.026 | 12234 |
| par contre | 18 | 12 | 0.72 | 0.989 | 0.044 | 3014 |
| par degrés | NON | NON | NON | NON | NON | 1447 |
| par dessus tout (i) | 3 | 6 | 0.91 | 0.988 | 0.075 | 1433 |
| par dessus tout (ii) | 2 | 10 | 0.73 | 0.991 | 0.028 | 1433 |
| pareil à | 14 | 10 | 0.68 | 0.983 | 0.041 | 6787 |
| par excellence | NON | NON | NON | NON | NON | 1749 |
| par faute de | 4 | 7 | 0.91 | 0.983 | 0.107 | 353 |
| parfois | 12 | 17 | 0.56 | 0.983 | 0.030 | 39445 |
| par hasard | 5 | 8 | 1.20 | 0.994 | 0.023 | 7071 |
| par instants | 0 | 9 | 0.73 | 0.993 | 0.041 | 1357 |
| par mégarde | NON | NON | NON | NON | NON | 578 |
| parmi d'autres | 9 | 12 | 0.86 | 0.991 | 0.030 | 620 |
| par moments | 0 | 12 | 0.88 | 0.984 | 0.031 | 2774 |
| par rapport à (i) | 0 | 9 | 0.91 | 0.985 | 0.043 | 5290 |
| par rapport à (ii) | 3 | 8 | 0.74 | 0.993 | 0.044 | 5290 |
| par surcrot | 19 | 12 | 0.59 | 0.982 | 0.034 | 498 |
| particulièrement | 14 | 18 | 0.51 | 0.982 | 0.038 | 12784 |


| Researched form | Latency | Growth | Slope | $\mathrm{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| par voie de | NON | NON | NON | NON | NON | 976 |
| par voie de conséquence | 0 | 9 | 0.56 | 0.983 | 0.058 | 130 |
| petit à petit | 3 | 10 | 0.77 | 0.985 | 0.030 | 1547 |
| peu à peu (i) | 10 | 6 | 1.83 | 0.997 | 0.014 | 16450 |
| peu à peu (ii) | 1 | 9 | 0.96 | 0.981 | 0.049 | 16450 |
| peu s'en faut | 0 | 9 | 0.92 | 0.982 | 0.030 | 221 |
| pour ainsi dire | 1 | 13 | 0.77 | 0.982 | 0.038 | 7704 |
| pour autant | 0 | 13 | 0.79 | 0.994 | 0.015 | 457 |
| pour finir | 23 | 15 | 0.63 | 0.982 | 0.039 | 838 |
| pour le coup | 14 | 6 | 0.72 | 0.986 | 0.123 | 464 |
| pour l'essentiel | 0 | 9 | 0.97 | 0.985 | 0.030 | 284 |
| pour le moment | 1 | 21 | 0.44 | 0.981 | 0.034 | 2986 |
| pour l'heure | NON | NON | NON | NON | NON | 546 |
| pour l'instant | 11 | 14 | 0.63 | 0.988 | 0.027 | 1859 |
| pour ma part | 5 | 6 | 1.12 | 0.997 | 0.097 | 2744 |
| pour peu que | 0 | 9 | 0.98 | 0.990 | 0.026 | 2479 |
| pour surcrot de | NON | NON | NON | NON | NON | 90 |
| pourtant que (i) | 0 | 6 | 1.05 | 0.989 | 0.054 | 4220 |
| pourtant que (ii) | 0 | 6 | 1.70 | 0.989 | 0.024 | 4220 |
| pour tout dire | 2 | 7 | 1.14 | 0.986 | 0.040 | 655 |
| pour un temps | NON | NON | NON | NON | NON | 1333 |
| présentement | 11 | 6 | 0.88 | 0.981 | 0.070 | 2683 |
| probablement (i) | 2 | 8 | 1.22 | 0.981 | 0.054 | 8497 |
| probablement (ii) | 2 | 10 | 0.70 | 0.981 | 0.033 | 8497 |
| proprement | 4 | 6 | 1.05 | 0.989 | 0.044 | 9817 |
| principalement | 17 | 7 | 1.31 | 0.993 | 0.029 | 6695 |
| progressivement | NON | NON | NON | NON | NON | 2235 |
| quand même | 3 | 14 | 0.57 | 0.993 | 0.019 | 12171 |
| quant à | 4 | 11 | 0.61 | 0.989 | 0.036 | 20878 |
| quant à cela | NON | NON | NON | NON | NON | 91 |
| quant à moi | 4 | 6 | 0.71 | 0.989 | 0.093 | 4875 |
| que dalle | NON | NON | NON | NON | NON | 163 |
| quelquefois | 13 | 7 | 1.44 | 0.982 | 0.033 | 34408 |
| quelque part | NON | NON | NON | NON | NON | 6454 |
| relatif à | 12 | 10 | 0.57 | 0.980 | 0.081 | 2850 |
| relativement à | 15 | 7 | 0.54 | 0.992 | 0.119 | 1469 |
| rien de plus | NON | NON | NON | NON | NON | 1537 |
| sans ambages | NON | NON | NON | NON | NON | 130 |
| sans commune mesure | 2 | 9 | 1.12 | 0.986 | 0.032 | 112 |
| sans crier gare | 0 | 11 | 0.69 | 0.989 | 0.026 | 211 |
| sans détour | 9 | 12 | 0.47 | 0.988 | 0.037 | 467 |
| sans façon | NON | NON | NON | NON | NON | 650 |
| sans tenir compte de | 5 | 8 | 0.85 | 0.983 | 0.037 | 143 |
| sauf | 4 | 8 | 1.03 | 0.983 | 0.038 | 11138 |
| sauf si | 3 | 12 | 0.74 | 0.994 | 0.027 | 247 |
| sauf que | 0 | 13 | 0.58 | 0.985 | 0.026 | 910 |
| selon moi | 1 | 13 | 0.39 | 0.982 | 0.066 | 1055 |
| si besoin est (i) | 5 | 6 | 2.30 | 0.992 | 0.035 | 106 |
| si besoin est (ii) | 3 | 6 | 2.27 | 0.995 | 0.021 | 106 |
| si bien que | 4 | 9 | 0.59 | 0.981 | 0.048 | 4831 |
| si ça se trouve | 0 | 7 | 1.21 | 0.986 | 0.029 | 144 |
| s'il en est | NON | NON | NON | NON | NON | 88 |
| si possible | 22 | 10 | 0.88 | 0.983 | 0.033 | 760 |
| soit dit en passant | NON | NON | NON | NON | NON | 276 |
| soudain | 28 | 12 | 0.63 | 0.989 | 0.031 | 3498 |
| soudainement | 0 | 6 | 2.38 | 0.980 | 0.044 | 94 |
| sous peu | 0 | 6 | 1.83 | 0.993 | 0.028 | 291 |
| sous prétexte de | NON | NON | NON | NON | NON | 2341 |
| sous prétexte que | 6 | 6 | 0.69 | 0.997 | 0.112 | 1364 |


| Researched form | Latency | Growth | Slope | $\mathrm{r}^{2}$ | Deviation | \# of occ. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| sous réserve que | NON | NON | NON | NON | NON | 89 |
| souventes fois | NON | NON | NON | NON | NON | 530 |
| spécialement | NON | NON | NON | NON | NON | 3764 |
| sur ce thème | NON | NON | NON | NON | NON | 130 |
| sur le champ | NON | NON | NON | NON | NON | 5152 |
| sur le moment | 8 | 16 | 0.38 | 0.992 | 0.033 | 715 |
| sur le sujet de | NON | NON | NON | NON | NON | 292 |
| sur le point de | 2 | 8 | 0.92 | 0.984 | 0.031 | 3321 |
| sur l'heure | NON | NON | NON | NON | NON | 720 |
| sur l'instant | 9 | 14 | 0.53 | 0.987 | 0.044 | 162 |
| un de ces jours | 5 | 6 | 1.31 | 0.996 | 0.026 | 983 |
| une sorte de | 1 | 6 | 1.82 | 0.985 | 0.034 | 31306 |
| une sorte de | 2 | 8 | 1.29 | 0.991 | 0.032 | 31306 |
| tandis que | BUG | BUG | BUG | BUG | BUG | 39303 |
| tant et plus | NON | NON | NON | NON | NON | 155 |
| tel quel | 4 | 8 | 1.07 | 0.983 | 0.036 | 985 |
| tour à tour | 11 | 22 | 0.37 | 0.982 | 0.043 | 4480 |
| tout à coup | 3 | 9 | 1.11 | 0.983 | 0.046 | 20468 |
| tout à fait | 25 | 12 | 0.46 | 0.981 | 0.047 | 25611 |
| tout à l'heure (i) | 7 | 9 | 0.99 | 0.985 | 0.040 | 12853 |
| tout à l'heure (ii) | 3 | 9 | 0.88 | 0.995 | 0.022 | 12853 |
| tout au long de | 13 | 12 | 0.68 | 0.980 | 0.035 | 1363 |
| tout au plus | 11 | 10 | 0.65 | 0.990 | 0.060 | 2954 |
| tout bien considéré | 2 | 6 | 1.16 | 0.994 | 0.038 | 152 |
| tout bien réfléchi | NON | NON | NON | NON | NON | 17 |
| tout compte fait | 7 | 6 | 0.86 | 0.995 | 0.068 | 390 |
| tout court | 0 | 6 | 2.49 | 0.985 | 0.046 | 1149 |
| tout de même | 34 | 13 | 0.74 | 0.991 | 0.017 | 13315 |
| tout du long | NON | NON | NON | NON | NON | 302 |
| toutefois | NON | NON | NON | NON | NON | 20576 |
| tout juste | 17 | 13 | 0.45 | 0.982 | 0.038 | 2055 |
| tout juste de | 4 | 6 | 0.60 | 0.982 | 0.107 | 197 |
| tout plein de | NON | NON | NON | NON | NON | 1014 |
| tout sauf | 12 | 9 | 0.46 | 0.995 | 0.075 | 158 |
| tout spécialement | 4 | 7 | 0.96 | 0.997 | 0.031 | 164 |
| tout un chacun | NON | NON | NON | NON | NON | 260 |
| très très | 3 | 14 | 0.51 | 0.991 | 0.081 | 356 |
| une espèce de | 8 | 6 | 1.87 | 0.981 | 0.046 | 12365 |
| un lendemain | 8 | 8 | 0.56 | 0.993 | 0.072 | 505 |
| un petit peu | 8 | 14 | 0.60 | 0.983 | 0.048 | 692 |
| un surcroît de | NON | NON | NON | NON | NON | 454 |
| un tas de | 23 | 9 | 1.30 | 0.990 | 0.023 | 4352 |
| venir de | 21 | 15 | 0.38 | 0.984 | 0.056 | 35884 |
| vis à vis de | 2 | 12 | 0.49 | 0.982 | 0.052 | 3384 |
| voilà | 3 | 13 | 0.75 | 0.984 | 0.033 | 90090 |
| vu que | 0 | 10 | 0.93 | 0.981 | 0.027 | 1230 |
| zut | 0 | 10 | 0.99 | 0.987 | 0.037 | 525 |

