CONSTRUCTING LANGUAGES TO EXPLORE THEORETICAL PRINCIPLES

Guillaume ENGUEHARD, *Xiaoliang* LUO, *Nicola* LAMPITELLI

ABSTRACT • The construction of languages has always been related to linguistics. It addresses linguistic notions or relevant questions from a non-academic point of view. In this paper, we propose a method inspired by experimental archaeology. The experiment consists in trying to obtain an artefact similar to the one observed using this or that construction method. An equivalent approach in linguistics would be the generation of linguistic systems based on explicitly formulated principles. Trying to generate similar systems pushes the linguist to explicitly define the principles that are needed and to explore all their consequences. In this context, we show that the use of notions induced by the observation of natural languages leads to a certain degree of circularity and that it is therefore more interesting to explore *a priori* principles based on very general assumptions.

KEYWORDS • Natural Languages; Natlang; Linguistic Theory; Constructed Languages; Conlangs.

1. Introduction

This paper aims to introduce an original approach to formal linguistics, which could be called 'simulationist linguistics', and which basically relies on two objectives:

1. to define a set of *a priori* principles concerning language.
2. to test the plausibility of those principles in constructing languages.

This approach differs strongly from the classical approach which consists in observing particular phenomena to induce general principles. In contrast, our goal is to consciously abstract as much as possible from particular phenomena in order to build a model whose value is strictly determined by its ability to generate data close to real language data, not to its ability to offer a good description of a particular language, to be realistic with regards to acquisition or to mimic cognitive mechanisms. Thus, the key word here is '*a priori* principles'. No matter these principles are of a structuralist, functionalist or generativist nature. The only thing that guides their selection is, again, their ability to produce naturalistic facts out of a minimum of assumptions. The present paper does not even try to define these principles, but to show the interest of the method.

In our first section, we show how constructions and discoveries are interrelated. We specifically focus on a parallel between experimental archeology and language construction. Second, we address important contrasts between modeling and simulation. We show that the former necessarily suffers from an inherent degree of circularity that only the second approach can overcome. Finally, we will discuss some examples of *a priori* principles for deriving phonological data in our third and fourth sections.

1. Construction and discovery

1.1. Constructed languages and linguistic research

The construction of languages has always been related to linguistics. Most of these initiatives address real scientific questions but from a non-academic point of view. The fact that Ferdinand de Saussure's own brother, René de Saussure, wrote a theoretical essay on the morphology of the most famous constructed language, esperanto, is an amusing illustration of this relation (de Saussure, 1914).

Constructing languages may be seen as a way to experiment or to exploit linguistic notions. Esperanto, is in itself an application of theoretical principles in order to facilitate mutual understanding: Zamenhof (1887) based his work on phonemic principle, paradigm uniformity, agglutination.

In the same way, Ogden's Basic English – a controlled language proposed as an alternative to Esperanto – is founded on an interesting and profound consideration for semantic primes (Ogden, 1930). Though some aspects of Ogden's work are criticized for their lack of objectivity or motivation, they have proven to be effective in the field of English language teaching.

Some constructed languages have fewer practical purposes. For instance, Loglan was created by Brown (1960) in order to test the Sapir-Worth hypothesis and Toki Pona was designed by Lang (2014) to take full advantage of the possibilities of polysemy in a minimal language.

1.2. Experimental archaeology

Experimental approaches based on artificial constructions have been explored in other fields, namely when:

1. there is no way to physically observe the cause of a given phenomenon.
2. the cause can hardly be deduced from the observation of the result.

In this case, causes are investigated through speculative methods. In archaeology, for instance, the causes of an artifact are regularly outside our empirical field and it is not possible to make abstract generalizations as we do in linguistics. Researchers therefore sometimes use speculation and validation of hypotheses through direct experience (Bordes, 1947). The experiment consists in seeing if one obtains an artifact similar to the one observed using this or that construction method. One key contribution from experimental archeology is to understand how items were produced in prehistoric times, especially in the lithic industry. We cannot observe the process of carving and the result of this process offers only very partial indications of their production method. The only way to solve the mystery of their origin is to reproduce the process through trial and error. Though it never gives definite answers, it makes it possible to (in)validate hypotheses on the scale of a complete system.

1.3. Experimentation and simulation

An equivalent approach applied to linguistics would be the generation of linguistic systems based on explicitly formulated principles. This approach is not strictly speaking 'experimental' linguistics. In linguistics, experimental methods are often intended to produce speech by controlling the conditions of its occurrence. The experiment is not in itself a hypothetical cause being tested, but an exclusion of confounding variables (Gillioz and Zufferey, 2020, p. 8). Thus, it aims to produce data that are difficult to observe in a natural context, but the conclusion refers to something broader than the data themselves or the conditions of the experiment. In contrast, experimental archeology aims to test a hypothesis that directly refers to the conditions of the experiment. We can therefore speak of indirect experience in the first case and direct experience in the second, which can be illustrated as follows:

1. Experimental linguistics: if A involves B, then C – which contains A – is true (indirect experience)
2. Experimental archeology: if A involves B, then A is true (direct experience)

Since experimental linguistics is not the exact equivalent of experimental archeology, we will speak here of 'simulationist linguistics' to refer to direct experience. We aim at testing directly potential causes of language systems. Our goal is to manipulate directly those parameters leading to a linguistic grammatical form rather than inferring them.

The origin of these parameters does not really matter. They may be the result of data observation, cognitive, behavioral, functional or even *a priori* speculation. Their value is ultimately determined only by the degree of closeness between the results obtained and the natural languages as a whole. Trying to generate similar systems pushes the linguist to explicitly define the principles that are needed and to explore all the consequences of these principles.

In the remainder of this paper, we are going to show how to reproduce a linguistic system and to make discoveries without inference.

2. Modeling and simulation

2.1. Language generators

There are many online generators for producing artificial linguistic data. Language generators such as Vulgarlang (https://www.vulgarlang.com/) stem from well-established observations of real linguistic facts. They thus manipulate *a posteriori* parameters: They combine current linguistic categories such as onsets, nuclei, codas, syllables, feet and words that result from a careful typological observation.

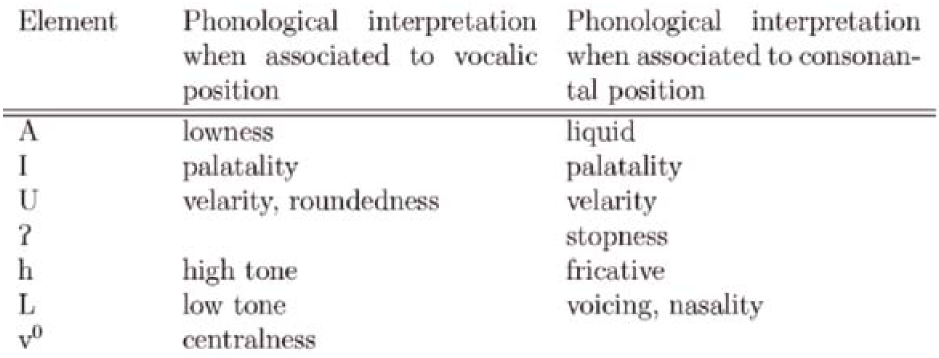
It is also the way in which current formal theories proceed. A presumably universal model is constructed from natural data, which is then declined according to parameters that do not seem to be reducible to a universal principle (Prince and Smolensky, 1993). This is what we call modeling.

2.2. Modeling vs Simulation

2.2.1. Modeling

What we call modeling is a two-step process. First, formal linguistics builds categories induced from the observation of natural languages in a bottom-up way, then it deduces possible linguistic systems through the application of such categories in a top-down way.

In phonology, for instance, Element Theory – henceforth ET – (Kaye et al., 1985; Backley, 2011) claims that sound inventory of any language is built from a limited number of universal, primitive units called Elements, these are induced from the observation of a large range of natural languages. A simplified version of the Elements is shown in Table 1.

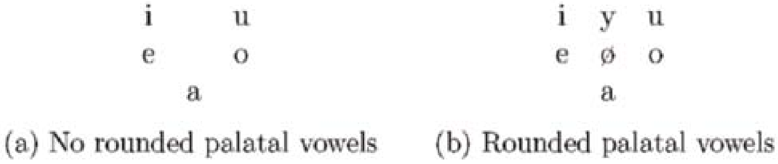


Element	Phonological interpretation when associated to vocalic position	Phonological interpretation when associated to consonantal position
A	lowness	liquid
I	palatality	palatality
U	velarity, roundedness	velarity
ʔ		stopness
h	high tone	fricative
L	low tone	voicing, nasality
v^0	centralness	

Table 1: Element Theory

Elements are then combined according to rules in order to derived balanced vowel systems. For instance, a mid-vowel [e] is the result of the following combination: AI.

We can take ET as a starting point to construct a language. A balanced vowel or consonant system can be considered as a structured application of combination rules, with a random variable handling the parameters of these combination rules. Table 2 shows two plausible vowel systems generated in a spreadsheet.



i		u
e		o
	a	

(a) No rounded palatal vowels

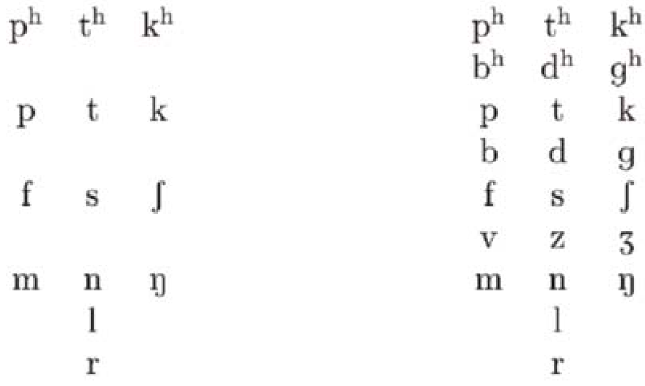
i	y	u
e	ø	o
	a	

(b) Rounded palatal vowels

Table 2: Example of generated vowel systems

Without going into details of the calculation of the functions in the spreadsheet, let us take 2a as an example. Table 2a allows the combination of I or U with A, resulting in three degrees of aperture. However, I and U cannot combine. In 2b, in contrast, the combination of I and U is allowed, which gives /y/ and /ø/ in the inventory.

The same principles can be applied to consonants. In Table 3 are shown two plausible consonant systems.



p^h	t^h	k^h
p	t	k
f	s	ʃ
m	n	ŋ
	l	
	r	

p^h	t^h	k^h
b^h	d^h	g^h
p	t	k
b	d	g
f	s	ʃ
v	z	ʒ
m	n	ŋ
	l	
	r	

Table 3: Example of generated consonant systems

In 3a, the marked consonant series is the aspirated one, there is no voiced consonant. In 3b, plosives can be marked by voicing, aspirated and both. The contrast between the two systems is due to the ability of L to combine with the element h found in obstruents.

As we can see in the above examples, one necessarily expects results derived from ET to be in conformity with natural languages since the former is based on generalizations from the latter. Generating constructed languages through modeling is thus efficient but tells us nothing new because of this circularity.

Moreover, *a posteriori* principles are partially biased because they are based on a limited sample of existing languages. Some languages are well described while others are not. These principles also hardly take into account areal explanations, and they cannot access prehistorical or future languages.

By the concept of 'simulation', we propose the reverse path: we define *a priori* principles from which we generate categories like onsets, nuclei, codas, and everything that makes linguistic data look natural.

2.2.2. Simulation

In contrast to empirical principles based on a sample of languages, *a priori* principles do not rely on the observation of languages. Their motivation can be diverse and ultimately matter very little. Only their ability to produce realistic data matters. By definition, such principles are universal and thus avoid the aforementioned biases.

Moreover, as modeling follows principles that are *a posteriori* derived from natural languages, generating a possible language on such a basis does nothing more than repeat an already known generalization. In the examples of Subsection 2.2.1, we managed to derive balanced systems with parametric variations only because we first assumed that phonological systems are balanced and subject to some specific parametric variations already implemented in our theory.

Conversely, if we can generate a possible language without assumptions drawn from natural languages, then it can tell us something new about the hidden mechanisms underlying the structure of natural languages. This is what we call 'simulation', as opposed to 'modeling'.

2.3. How to define language naturalness?

It is difficult to define exactly what is natural in a language. To have the answer to this question, we need to know what is possible and what is impossible, and thus to have an exact model of how languages work.

In the meantime, we propose a minimalist definition according to which the naturalness of languages is a set of parametric restrictions affecting the following aspects of languages:

1. The inventory of units (= what)
2. The qualitative combination of units (= how)
3. The quantitative combination of units (= how much)

3. Modules and derivation

In this section, we present three different modules dedicated to deriving the phonological inventory, syllabic constraints, and word size respectively. These modules were implemented in a simple spreadsheet (LibreOffice Calc) with formula that everyone can reproduce.[1] In each case, we first present the *a priori* principle on which the module is based, and then we illustrate our results using a chosen simulation example as well as a randomly selected simulation example.

All the principles retained in our presentation remain debatable. They only serve to illustrate our point about the language simulation research method.

3.1. Module 1: Inventory

3.1.1. Assumption and implementation

Our first module is based on a functionalist assumption that the members of an inventory should be as far apart as possible in order to maximize the contrast between the different phonemes. This principle recalls Martinet's diachronic phonology (Martinet, 1955) and the Adaptive Dispersion Theory (Liljencrantz and Lindblom, 1972).

In practice, this postulate imposes to define a continuous space of phonetic values from which the inventory is built by selection. We define such a space in Table 4. For instance, this table does not show any dichotomy between consonants and vowels. Of course, this example is far from satisfactory. It is limited by a two-dimensional representation that does not account for the proximity between labials and gutturals, and the use of 'weird' symbols only loosely simulates a continuous space. For these two reasons, it is futile to present our category choices in more detail. But this example is sufficient, for the moment, to illustrate our point.

[1] The main formula can be found in the appendix.

p	p̪	t̪	t	c͡p	k͡p	t̠	ʈ	c	k	q	ʔ
b	b̪	d̪	d	ɟ͡b	ɡ͡b	d̠	ɖ	ɟ	ɡ	ɢ	
ɸ	f	θ	s	ç͡ɸ	x͡ɸ	ʃ	ʂ	ç	x	χ	h
β	v	ð	z	ʝ͡β	ɣ͡β	ʒ	ʐ	ʝ	ɣ	ʁ	ɦ
m	ɱ	n̪	n	m͡ɲ	m͡ŋ	n̠	ɳ	ɲ	ŋ	ɴ	
		l̪	l			l̠	ɭ	ʎ	ʟ		
		r̪	r			r̠	ɽ͡r			ʀ	
		ɾ̪	ɾ			ɾ̠	ɽ				
	ʋ	ð̞	ɹ	ɥ	w	ɹ̠	ɻ	j	ɰ	ɰ̠	ɨ̞
				y	u	i̠	ɨ̠	i	ɯ	ɯ̠	ɨ
				ʏ	ʊ	ɪ	ï̠	ɪ̠	ɯ̆	ɯ̠̆	ï
				ø	o	e̠	ɘ̠	e	ɤ	ɤ̠	ɘ
				œ	ɔ	ɛ̠	ɜ̠	ɛ	ʌ	ʌ̠	ɜ
				ɶ	ɒ	æ̠	a̠	æ	ɑ	ɑ̠	a

Table 4: Module 1 – Suggestion for a phonic continuum

The definition of the phonological inventory is done by applying a categorial mesh to this space. This mesh follows the principles mentioned above, i.e. the fact that it exploits the entire phonological continuum and seeks to maximize the distance between its nodes.

In order to allow some variation in the realization of these principles, we introduce two random parameters following a normal distribution law: the first one manages the tightness of the mesh, as in Figures 1a and 1b; and the second one manages the regularity of its nodes distribution, as in Figures 1b and 1c.

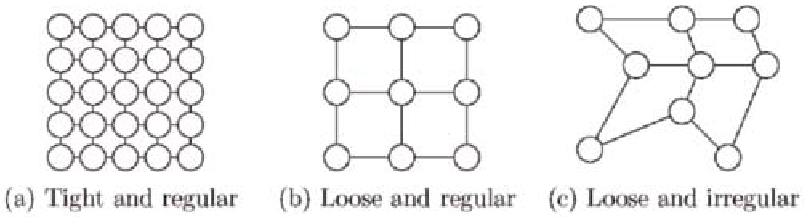

(a) Tight and regular (b) Loose and regular (c) Loose and irregular

Figure 1: Mesh types

Each phonetic value falling under a node of the mesh is then selected in the phonological inventory.

3.1.2. Results

In the chosen example in Table 5, the implementation of our assumption simulates a system with 7 obstruants, 4 sonorants and 3 vowels. To the reader put off by the presence of unusual sounds, note that the selected symbols still represent specific phonetic values, not phonemes gathering a whole range of phonetic values.

A more phonological reading of this system reveals a system with three cardinal vowels /i,a,u/.

p	t̼			ʈ	q
β		j͡β	ʒ		
					ɴ
	ɾ̼			ɽ	
				ï̠	ɯ̠̆
				a̠	

Table 5: Module 1 – Chosen example (after 9 generations)

In the random example in Table 6, the vowel system is typologically distinct: it is more akin to a 5-vowel system like /i,e,a,o,u/. But in both cases, we observe universal patterns such as the presence of obstruents, sonorants and vowels, or the existence of several place features.

p		t̼		k͡p		ʈ		q
					ʃ		x	
	v							
			l			ɭ		
		ð̩̆				ɻ		ɰ̠
				ʊ		ï̠		ɯ̠̆
						ɜ̠	ʌ	
				ɒ				

Table 6: Module 1 – Random example

Thus, a very simple *a priori* principle succeeds in simulating various types of phonological systems without help of induced typological conjectures.

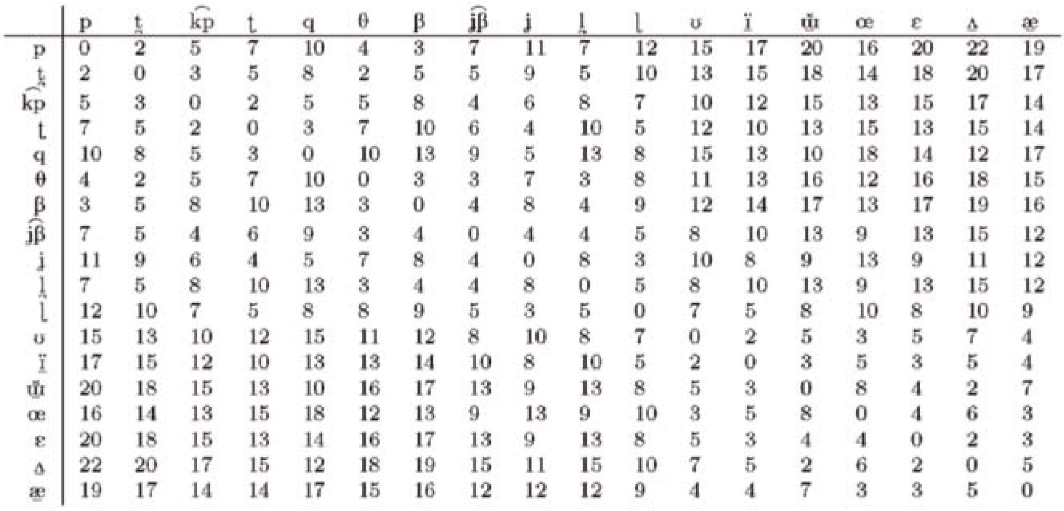
3.2. Module 2: Qualitative combination

3.2.1. Assumption and implementation

Now that we have defined our inventory, we need to define the way all these values can combine. The use of concepts such as syllable, nucleus, onset, coda, foot, etc. is totally excluded, as they are based on an observation of facts.

Our *a priori* assumption for the second module is based on the same functionalist mechanism as above. Just as the units must be maximally contrastive on the paradigmatic axis, they must be maximally contrastive on the syntagmatic axis. This also recalls notions explored in Hjelmslev (1943) or OCP (Goldsmith, 1976; McCarthy, 1979). Thus, sequences of segments are ruled by the distance between two values in the phonetic space.

We calculate the distance between two values of a given inventory with tables like the one in 7. Each number represents the number of cells between two phonetic values.



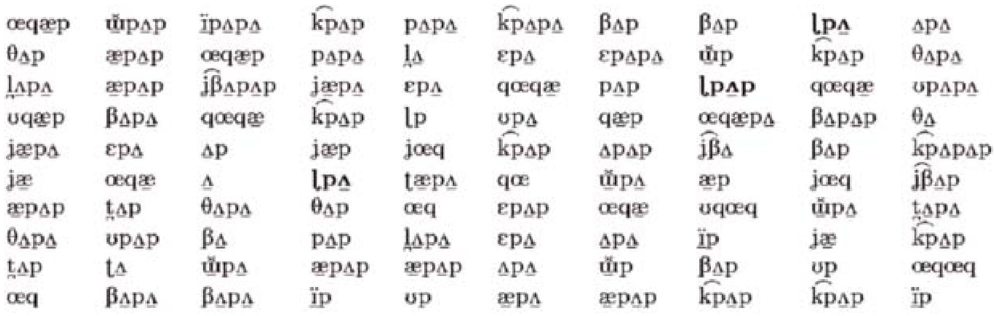
	p	t̪	k͡p	ʈ	q	θ	β	j͡β	j	l̪	ɭ	ʊ	ɪ̈	ɯ̆ɪ	œ	ɛ	ʌ	æ
p	0	2	5	7	10	4	3	7	11	7	12	15	17	20	16	20	22	19
t̪	2	0	3	5	8	2	5	5	9	5	10	13	15	18	14	18	20	17
k͡p	5	3	0	2	5	5	8	4	6	8	7	10	12	15	13	15	17	14
ʈ	7	5	2	0	3	7	10	6	4	10	5	12	10	13	15	13	15	14
q	10	8	5	3	0	10	13	9	5	13	8	15	13	10	18	14	12	17
θ	4	2	5	7	10	0	3	3	7	3	8	11	13	16	12	16	18	15
β	3	5	8	10	13	3	0	4	8	4	9	12	14	17	13	17	19	16
j͡β	7	5	4	6	9	3	4	0	4	4	5	8	10	13	9	13	15	12
j	11	9	6	4	5	7	8	4	0	8	3	10	8	9	13	9	11	12
l̪	7	5	8	10	13	3	4	4	8	0	5	8	10	13	9	13	15	12
ɭ	12	10	7	5	8	8	9	5	3	5	0	7	5	8	10	8	10	9
ʊ	15	13	10	12	15	11	12	8	10	8	7	0	2	5	3	5	7	4
ɪ̈	17	15	12	10	13	13	14	10	8	10	5	2	0	3	5	3	5	4
ɯ̆ɪ	20	18	15	13	10	16	17	13	9	13	8	5	3	0	8	4	2	7
œ	16	14	13	15	18	12	13	9	13	9	10	3	5	8	0	4	6	3
ɛ	20	18	15	13	14	16	17	13	9	13	8	5	3	4	4	0	2	3
ʌ	22	20	17	15	12	18	19	15	11	15	10	7	5	2	6	2	0	5
æ	19	17	14	14	17	15	16	12	12	12	9	4	4	7	3	3	5	0

Table 7: Module 2 – Proximity calculation

Then we apply a random parameter which, for each segment of a word, selects the next one among the most distant values following a normal distribution law.

3.2.2. Results

In the chosen example in Table 8, our implementation simulates pseudo-words with mostly CV, CVC, V, VC syllables. All the words contain a vowel. Only some of them, in bold, contain a consonant cluster, which is always [ɭp].

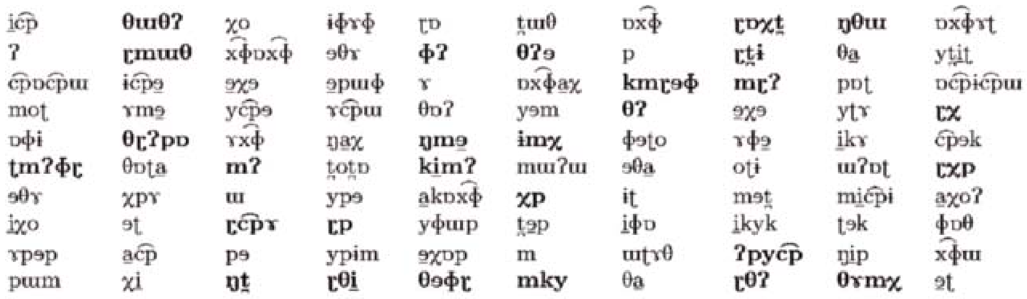


œqæp	ɯ̆ɪpʌp	ɪ̈pʌpʌ	k͡pʌp	pʌpʌ	k͡pʌpʌ	βʌp	βʌp	**ɭpʌ**	ʌpʌ
θʌp	æpʌp	œqæp	pʌpʌ	l̪ʌ	ɛpʌ	ɛpʌpʌ	ɯ̆ɪp	k͡pʌp	θʌpʌ
l̪ʌpʌ	æpʌp	j͡βʌpʌp	jæpʌ	ɛpʌ	qœqæ	pʌp	**ɭpʌp**	qœqæ	ʊpʌpʌ
ʊqæp	βʌpʌ	qœqæ	k͡pʌp	ɭp	ʊpʌ	qæp	œqæpʌ	βʌpʌp	θʌ
jæpʌ	ɛpʌ	ʌp	jæp	jœq	k͡pʌp	ʌpʌp	j͡βʌ	βʌp	k͡pʌpʌp
jæ	œqæ	ʌ	**ɭpʌ**	ʈæpʌ	qœ	ɯ̆ɪpʌ	æp	jœq	j͡βʌp
æpʌp	t̪ʌp	θʌpʌ	θʌp	œq	ɛpʌp	œqæ	ʊqœq	ɯ̆ɪpʌ	t̪ʌpʌ
θʌpʌ	ʊpʌp	βʌ	pʌp	l̪ʌpʌ	ɛpʌ	ʌpʌ	ɪ̈p	jæ	k͡pʌp
t̪ʌp	ʈʌ	ɯ̆ɪpʌ	æpʌp	æpʌp	ʌpʌ	ɯ̆ɪp	βʌp	ʊp	œqœq
œq	βʌpʌ	βʌpʌ	ɪ̈p	ʊp	æpʌ	æpʌp	k͡pʌp	k͡pʌp	ɪ̈p

Table 8: Module 2 – Chosen example (after 8 generations)

In the random example in Table 9, the simulated words are more complex. Many of them, in bold, contains consonant clusters, and some have no vowel. This absence of vowel can be intimidating, as in [ʈmʔɸɽ], but it should be noted that most words contain a vowel.

Our implementing system says nothing about consonant syllabicity, but we could admit that it is derived from a consonant surrounded by more obstruent consonants. In this case, the following pseudo-language resembles Berber.



i̯c͡p	**θɯθʔ**	χo	**iɸɤɸ**	ɽɒ	t̪ɯθ	ɒx͡ɸ	**ɽɒχt̪**	**ŋθɯ**	ɒx͡ɸɤʈ
ʔ	**ɽmɯθ**	x͡ɸɒx͡ɸ	əθɤ	**ɸʔ**	**θʔə**	p	**ɽt̪ɨ**	θa̰	yt̪i̯ʈ
c͡pɒc͡pɯ	ɨc͡pə̰	ə̰χə	ə̰pɯɸ	ɤ	ɒx͡ɸa̰χ	**kmɽəɸ**	**mɽʔ**	pɒʈ	ɒc͡pɨc͡pɯ
moʈ	ɤmə̰	yc͡pə	ɤc͡pɯ	θɒʔ	yəm	**θʔ**	ə̰χə	yʈɤ	**ɽχ**
ɒɸi	**θɽʔpɒ**	ɤx͡ɸ	ŋa̰χ	**ŋmə̰**	**ɨmχ**	ɸəʈo	ɤɸə̰	i̯kɤ	c͡pək
ʈmʔɸɽ	θɒʈa	**mʔ**	t̪ot̪ɒ	**ki̯mʔ**	mɯʔɯ	əθa	oʈɨ	ɯʔɒʈ	**ɽχp**
əθɤ	χpɤ	ɯ	ypə	a̰kɒx͡ɸ	**χp**	iʈ	mət̪	mi̯c͡pi	a̰χoʔ
i̯χo	əʈ	**ɽc͡pɤ**	**ɽp**	yɸɯp	t̪ə̰p	i̯ɸɒ	i̯kyk	ʈək	ɸɒθ
ɤpəp	a̰c͡p	pə	ypim	ə̰χɒp	m	ɯʈɤθ	**ʔpyc͡p**	ŋip	x͡ɸɯ
pɯm	χi̯	**ŋt̪**	**ɽθi̯**	**θəɸɽ**	**mky**	θa	**ɽθʔ**	**θɤmχ**	əʈ

Table 9: Module 2 – Random example

Once again, an *a priori* principle is sufficient to simulate different types of syllabic constraints without introducing the notion of syllable or without referring to the patterns of natural languages. The results obviously deserve to be improved, but they already show that we can explain a lot with very little.

3.3. Module 3: Quantitative combination

3.3.1. Assumption and implementation

Our last module aims to constrain the size of the simulated words. Still adopting a functionalist viewpoint, we assume that words must be as short as possible to preserve articulatory energy, but they must be large enough to allow the creation of a lexicon composed of distinct elements. In sum, the average word size must be strictly sufficient to allow a number of distinct combinations that corresponds to a universal lexicon size. This postulate is still in line with the logic of maximizing contrasts.

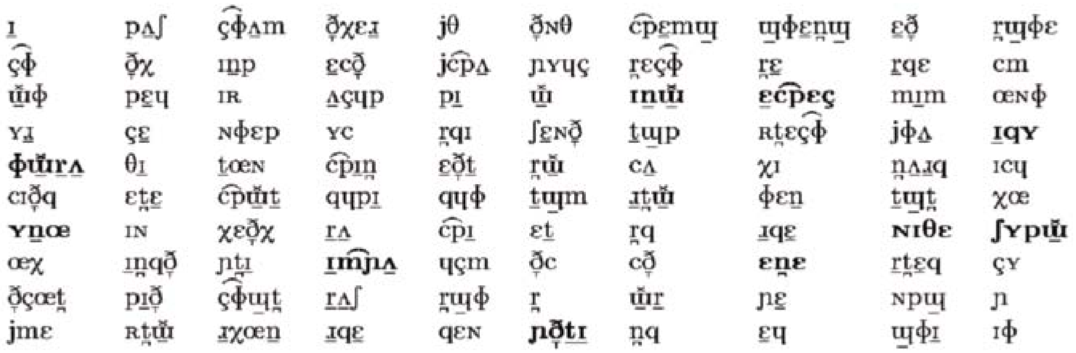
To implement this, we define a lexicon size arbitrarily (beyond a certain size, the difference doesn't make much difference) and we count the number of units in the phonological inventory. Then we calculate the average word size that is needed to derive a number of distinct combinations equal to the lexicon size, as in the example below:

Words: 200 000
Segments: 34
Average word size: 3,461379675

As before, we introduce the possibility of variation by applying a random parameter that defines the variation margin of each word from a normal distribution law.

3.3.2. Results

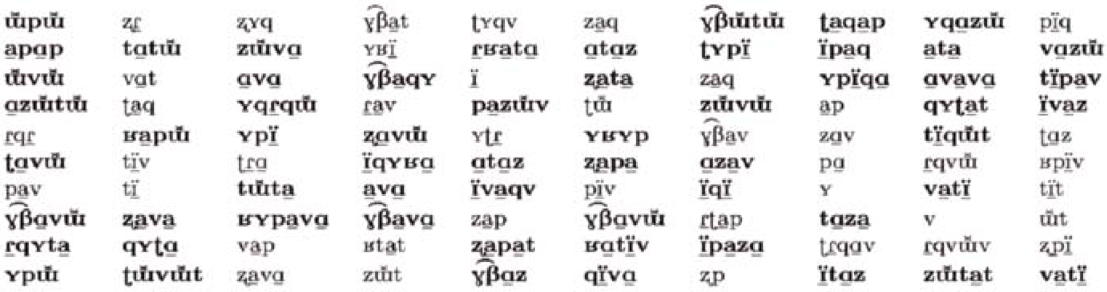
In the chosen example in Table 10, we obtain mostly 1-syllable words. Only a very small number of the pseudo-words, in bold, contain two syllables.



ɪ̠	pʌ̠ʃ	ç͡ɸʌ̠m	ð̞χɛɹ̠	jθ	ð̞ɴθ	c͡pɛ̠mɰ̠	ɰ̠ɸɛ̠n̪ɰ̠	ɛ̠ð̞	r̪ɰ̠ɸɛ
ç͡ɸ	ð̞χ	ɪn̠p	ɛ̠cð̞	jc͡pʌ̠	ɲʏɥç	r̪ɛç͡ɸ	r̪ɛ̠	r̠qɛ	cm
ɯ̽ɸ	pɛ̠ɥ	ɪʀ	ʌ̠çɥp	pɪ̠	ɯ̽	**ɪn̪ɯ̽**	**ɛ̠c͡pɛç**	mɪ̠m	œɴɸ
ʏɹ̠	çɛ̠	ɴɸɛp	ʏc	r̪qɪ	ʃɛ̠ɴð̞	t̪ɰ̠p	ʀt̪ɛç͡ɸ	jɸʌ̠	**ɪ̠qʏ**
ɸɯ̽r̠ʌ̠	θɪ̠	t̠œɴ	c͡pɪn̪	ɛ̠ð̞t̠	r̪ɯ̽	cʌ̠	χɪ	n̪ʌ̠ɹ̠q	ɪcɥ
cɪð̞q	ɛt̪ɛ̠	c͡pɯ̽t̠	qɥpɪ̠	qɥɸ	t̠ɰ̠m	ɹ̠t̪ɯ̽	ɸɛn̪	t̠ɰ̠t̪	χœ
ʏn̪œ	ɪɴ	χɛð̞χ	r̠ʌ	c͡pɪ̠	ɛt̠	r̪q	ɹ̠qɛ̠	**ɴɪθɛ**	**ʃʏpɯ̽**
œχ	ɪ̠n̪qð̞	ɲt̪ɪ	**ɪ̠m͡ɲɪʌ̠**	ɥçm	ð̞c	cð̞	**ɛn̪ɛ**	r̠t̪ɛ̠q	çʏ
ð̞çœt̪	pɪð̞	ç͡ɸɰ̠t̪	r̠ʌʃ	r̪ɰ̠ɸ	r̪	ɯ̽r̠	ɲɛ̠	ɴpɰ̠	ɲ
jmɛ	ʀt̪ɯ̽	ɹ̠χœn̪	ɹ̠qɛ̠	qɛɴ	**ɲð̞tɪ̠**	n̪q	ɛ̠q	ɰ̠ɸɪ̠	ɪɸ

Table 10: Module 3 – Chosen example (after 15 generations)

In the random example in Table 11, we can see that most simulated words have more than one syllable, and some of them can go up to three syllables, as [ɑz̠ɯ̽tɯ̽].



ɯ̽pɯ̽	z̡r̠	z̡ʏq	ɣ͡βa̠t	t̡ʏqv	za̠q	**ɣ͡βɯ̽tɯ̽**	**t̡a̠qa̠p**	**ʏqɑ̠zɯ̽**	pï̠q
a̠pɑp	**tɑtɯ̽**	**zɯ̽vɑ**	ʏʙï̠	**r̠ʙa̠tɑ**	**ɑtɑz**	**t̡ʏpï̠**	**ï̠pa̠q**	**a̠ta**	**vɑzɯ̽**
ɯ̽vɯ̽	vɑt	**ɑvɑ**	**ɣ͡βaqʏ**	ï	**z̡a̠ta**	za̠q	**ʏpï̠qɑ**	**ɑva̠vɑ**	**tï̠pa̠v**
ɑzɯ̽tɯ̽	t̡a̠q	**ʏqr̠qɯ̽**	r̠av	**pa̠zɯ̽v**	t̡ɯ̽	**zɯ̽vɯ̽**	a̠p	**qʏt̡a̠t**	**ï̠va̠z**
r̠qr̠	**ʙa̠pɯ̽**	**ʏpï̠**	**z̡ɑvɯ̽**	ʏt̡r̠	**ʏʙʏp**	ɣ͡βa̠v	zɑv	**tï̠qɯ̽t**	t̡ɑz
t̡ɑvɯ̽	tï̠v	t̡rɑ	**ï̠qʏʙɑ**	**ɑtɑz**	**z̡a̠pa̠**	**ɑza̠v**	pɑ	r̠qvɯ̽	ʙpï̠v
pa̠v	tï̠	**tɯ̽ta̠**	**a̠vɑ**	**ï̠va̠qv**	pï̠v	**ï̠qï̠**	ʏ	**va̠tï̠**	tï̠t
ɣ͡βɑvɯ̽	**z̡a̠va̠**	**ʙʏpa̠vɑ**	**ɣ͡βa̠vɑ**	za̠p	**ɣ͡βɑvɯ̽**	r̠t̡a̠p	**tɑza̠**	v	ɯ̽t
r̠qʏta̠	**qʏt̡ɑ**	va̠p	ʙta̠t	**z̡a̠pa̠t**	**ʙɑtï̠v**	**ï̠pa̠zɑ**	t̡r̠qɑv	r̠qvɯ̽v	z̡pï̠
ʏpɯ̽	**t̡ɯ̽vɯ̽t**	z̡a̠vɑ	zɯ̽t	**ɣ͡βɑz**	**qï̠vɑ**	z̡p	**ï̠tɑz**	**zɯ̽ta̠t**	**va̠tï̠**

Table 11: Module 3 – Random example

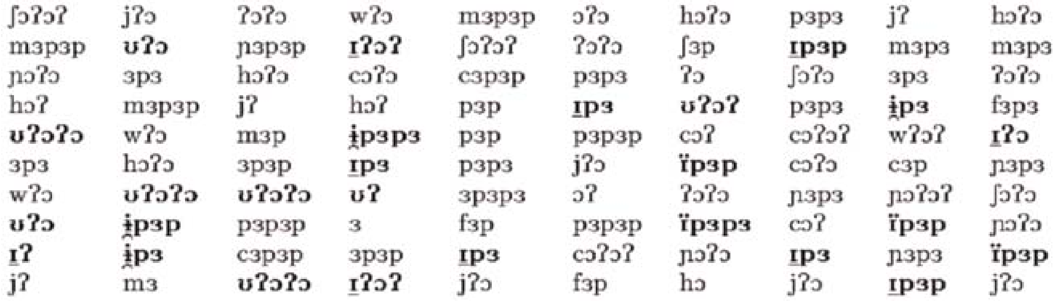
Thus, our *a priori* assumption can simulate languages with very various word sizes. Of course, morphology has an obvious role in defining the word size and a complete simulation must take it into account. But it is interesting to note that phonological principles may also be sufficient to explain the variation of languages on this issue.

4. Deriving a notion without specific module

We end our statement with an interesting case illustrating how a linguistic phenomenon can emerge without the help of a module designed specifically for this purpose.

In the six examples discussed above, a module of a priori principles was defined in order to derive a range of combinations under the same notion: phonological inventory, phonotactics or minimal word. The example below illustrates the derivation of a new notion, stress, without any new module being added.

This can be observed in Table 12 where closed vowels, in bold, can be found only in initial syllables.[2] The other vowels can be found in initial or non-initial syllable. This reduction of the inventory in some syllables is what we usually call a vowel reduction, and this indicates here the presence of stress in initial syllable.



ʃɔʔɔʔ	jʔɔ	ʔɔʔɔ	wʔɔ	mɜpɜp	ɔʔɔ	hɔʔɔ	pɜpɜ	jʔ	hɔʔɔ
mɜpɜp	**ʊʔɔ**	ɲɜpɜp	**ɪ̱ʔɔʔ**	ʃɔʔɔʔ	ʔɔʔɔ	ʃɜp	**ɪ̱pɜp**	mɜpɜ	mɜpɜ
ɲɔʔɔ	ɜpɜ	hɔʔɔ	cɔʔɔ	cɜpɜp	pɜpɜ	ʔɔ	ʃɔʔɔ	ɜpɜ	ʔɔʔɔ
hɔʔ	mɜpɜp	jʔ	hɔʔ	pɜp	**ɪ̱pɜ**	**ʊʔɔʔ**	pɜpɜ	**ɨ̱pɜ**	fɜpɜ
ʊʔɔʔɔ	wʔɔ	mɜp	**ɨ̱pɜpɜ**	pɜp	pɜpɜp	cɔʔ	cɔʔɔʔ	wʔɔʔ	**ɪ̱ʔɔ**
ɜpɜ	hɔʔɔ	ɜpɜp	**ɪ̱pɜ**	pɜpɜ	jʔɔ	**ïpɜp**	cɔʔɔ	cɜp	ɲɜpɜ
wʔɔ	**ʊʔɔʔɔ**	**ʊʔɔʔɔ**	**ʊʔ**	ɜpɜpɜ	ɔʔ	ʔɔʔɔ	ɲɜpɜ	ɲɔʔɔʔ	ʃɔʔɔ
ʊʔɔ	**ɨ̱pɜp**	pɜpɜp	ɜ	fɜp	pɜpɜp	**ïpɜpɜ**	cɔʔ	**ïpɜp**	ɲɔʔɔ
ɪ̱ʔ	**ɨ̱pɜ**	cɜpɜp	ɜpɜp	**ɪ̱pɜ**	cɔʔɔʔ	ɲɔʔɔ	**ɪ̱pɜ**	ɲɜpɜ	**ïpɜp**
jʔ	mɜ	**ʊʔɔʔɔ**	**ɪ̱ʔɔʔ**	jʔɔ	fɜp	hɔ	jʔɔ	**ɪ̱pɜp**	jʔɔ

Table 12: Example of initial stress and vowel reduction

This case is a key example of the research method we propose. It shows how the simulation can formally invalidate the primitiveness of a given notion.

Of course, this example represents only one of a range of possible combinations classified under the notion of stress, and one might object that the notion of stress is still necessary to account for all the associated phenomena. But to show that there is at least one case where the notion of stress is redundant is sufficient to question linguists who, being too used to manipulating this notion, never call it into question.

Conclusion

In this paper, we showed that modeling and simulation are two different things and that simulating plausible languages helps to understand what categories are primitive and what categories can be derived from more general principles.

Do our simulated languages look like natural languages? They do in the sense that our approach derives plausible languages displaying various types of restrictions on the inventory and the combination of units.

Our approach is very distinct from those based on modeling, which aim to define prototypical languages. Indeed, the aforementioned parametric restrictions do not emerge from categories induced by the observation of natural languages: they emerge from nothing but *a priori* principles.

The reader should keep in mind that the usual definition of a prototypical model is itself biased by the nature of the most observed languages, European languages. But many languages surprise us every day with their ‘unnatural’ aspects. Thus, though some of our results may seem odd or unnatural – and it is sure that they should be improved – this should not be used to reject without nuance this method. Experimental archaeologists faced similar difficulties in the early days of the discipline:

[2] It should be mentioned that these simulated words were obtained with a phonetic continuum slightly different from the one presented in this article.

> I don't claim to have the "orthodox" method of carving. It is very possible that the "Mousterian" technique exposed is not the one that was used by the Mousterians, or by all the Mousterians.
> These experiments are still in progress. They have not yet allowed me to completely reproduce the magnificent retouching "en echarpe" of Egyptian flints. (Bordes, 1947, p. 1-2)

The most important is that we can simulate some of the aspects found in natural languages with a minimum of assumptions. For instance, we found a pseudo-language with initial stress and vowel reduction without setting the category of stress. This implies to ask whether the notion of stress is ultimately necessary when we manage to simulate, without it, a process generally attributed to prosody.

Next, we should adapt our principles to morphology, syntax and semantics. We should also improve our results in phonology, for example by defining a more satisfying phonetic continuum. But for the time being, we showed to what extent constructing languages is an interesting way to do science.

REFERENCES

Backley, P. (2011). *An Introduction to Element Theory*. Edinburgh University Press.

Bordes, F. (1947). Etude comparative des différentes techniques de taille du silex et des roches dures. *L'anthropologie,* 51:1-29.

Brown, J. C. (1960). Loglan. *Scientific American,* 202(6):53-63.

de Saussure, R. (1914). *La vort-teorio en Esperanto*. Universala Esperantia Librejo.

Gillioz, C. and Zufferey, S. (2020). *Introduction to Experimental Linguistics*. ISTE, London.

Goldsmith, J. A. (1976). *Autosegmental Phonology*. PhD dissertation [ms], MIT.

Hjelmslev, L. (1943). *Prolegomena to a Theory of Language*. University of Wisconsin Press, Madison, [1969] edition.

Kaye, J., Lowenstamm, J., and Vergnaud, J.-R. (1985). The internal structure of phonological representations: a theory of charm and government. *Phonology Yearbook*, 2:305-328.

Lang, S. (2014). *Toki Pona: The Language of Good*. CreateSpace, Charleston.

Liljencrantz, J. and Lindblom, B. (1972). Numerical simulation of Vowel Quality Systems: The Role of Perceptual Contrasts. *Language*, 48(4):839- 862.

Martinet, A. (1955). *Economie des changements phonetiques: traité de phonologie diachronique*. A. Francke, Berne.

McCarthy, J. J. (1979). *Formal problems in Semitic phonology and morphology*. PhD dissertation [ms], MIT Cambridge, Mass.

Ogden, C. K. (1930). *Basic English: A General Introduction with Rules and Grammar*. Paul Treber & Co., London.

Prince, A. S. and Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers Center for Cognitive Science, New Brunswick, [2002] edition.

Zamenhof, L.-L. (1887). *Mezhdunarodnyj Jazyk*. Kh. Kel'ter, Varsovia.

GUILLAUME ENGUEHARD • is Assistant Professor of Linguistics at Université d'Orléans and the reseach lab LLL (UMR 7270 CNRS); he works in the field of theoretical phonology and has published articles on the role of time units and phonological contrast in stress and syllable structure. He is also working on defining the varieties of Old Norse formerly spoken in Normandy.

XIAOLIANG LUO • is Associate professor in Chinese linguistics at Université Paris-Cité and at the research lab Histoire des Théories Linguistiques (UMR 7597 of the CNRS). His research interests include theoretical phonology, Chinese phonology and history of linguistics. His recent work focuses on fortis and lenis on the one hand, melody and prosody on the other.

NICOLA LAMPITELLI • is Professor of Linguistics at Université Paris Nanterre and the research lab MoDyCo (UMR 7114 CNRS); he works on the phonology and the morphology of Romance languages (mainly Italian and its dialects, and French) and Afroasiatic languages (Somali, but also Neo-Aramaic and Arabic), and he recently published articles on *Vowel Length in Friulian verbs: a case of mora affixation* and *Conditions on complex exponence: a case study of the Somali subject marker*. His research interests also include field linguistics and language description.

Appendix: formula

Module 1

```
=(SIN((2*PI()/V!$C$3)*((ROW(A1)-2)-(V!$C$3/4)))+1)+(SIN((2*PI()/V!$C$4)*((COLUMN(A1)-2)-(V⌋
↪  !$C$4/4)))+1)+SQRT(NORM.INV(RAND();0;25)^2)*(SIN((2*PI()/V!$C$3)*((ROW(A1)-2)-(V!$C$3/⌋
↪  4)))+1)+(SIN((2*PI()/V!$C$4)*((COLUMN(A1)-2)-(V!$C$4/4)))+1)/100
```

```
V!$C$3=ArrayFormula(LARGE(IF(P!$A$1:$Z$1000<>"";COLUMN(P!$A$1:$Z$1000));1))/((50+NORM.INV(⌋
↪  RAND();0;12,5))*ArrayFormula(LARGE(IF(P!$A$1:$Z$1000<>"";COLUMN(P!$A$1:$Z$1000));1))/1⌋
↪  00)
```

```
V!$C$4=ArrayFormula(LARGE(IF(P!$A$1:$Z$1000<>"";ROW(P!$A$1:$Z$1000));1))/((50+NORM.INV(RAN⌋
↪  D();0;12,5))*ArrayFormula(LARGE(IF(P!$A$1:$Z$1000<>"";ROW(P!$A$1:$Z$1000));1))/100)
```

```
P!$A$1:$Z$1000= [phonetic space]
```

Modules 2 and 3

```
=IF(B1="";"";IF(COLUMN(C1)-1>V!$B$6+INT(NORM.INV(RAND();0;25))*V!$B$6/100;"";DECALER(INDIR⌋
↪  ECT(ADRESS(MATCH(B1;T!$A:$A;0);ArrayFormula(SMALL(IF(INDIRECT(ADRESS(MATCH(B1;T!$A:$A;⌋
↪  0);2;;;"T")):INDIRECT(ADRESS(MATCH(B1;T!$A:$A;0);ArrayFormula(SMALL(IF(T!$1:$1="";COLU⌋
↪  MN(T!$1:$1));2))-1;;;"T"))>=INT(INDIRECT(ADRESS(MATCH(B1;T!$A:$A;0);ArrayFormula(SMALL⌋
↪  (IF(T!$1:$1="MAX";COLUMN(T!$1:$1));1));;;"T"))-(V!$B$2*INDIRECT(ADRESS(MATCH(B1;T!$A:$⌋
↪  A;0);ArrayFormula(SMALL(IF(T!$1:$1="MAX";COLUMN(T!$1:$1));1));;;"T"))/100));COLUMN(IND⌋
↪  IRECT(ADRESS(MATCH(B1;T!$A:$A;0);2;;;"T")):INDIRECT(ADRESS(MATCH(B1;T!$A:$A;0);ArrayFo⌋
↪  rmula(SMALL(IF(T!$1:$1="";COLUMN(T!$1:$1));2))-1;;;"T"))));INT(RAND()*NB.IF(INDIRECT(A⌋
↪  DRESS(MATCH(B1;T!$A:$A;0);2;;;"T")):INDIRECT(ADRESS(MATCH(B1;T!$A:$A;0);ArrayFormula(S⌋
↪  MALL(IF(T!$1:$1="";COLUMN(T!$1:$1));2))-1;;;"T"));">="&INT(INDIRECT(ADRESS(MATCH(B1;T!⌋
↪  $A:$A;0);ArrayFormula(SMALL(IF(T!$1:$1="MAX";COLUMN(T!$1:$1));1));;;"T"))-(V!$B$2*INDI⌋
↪  RECT(ADRESS(MATCH(B1;T!$A:$A;0);ArrayFormula(SMALL(IF(T!$1:$1="MAX";COLUMN(T!$1:$1));1⌋
↪  ));;;"T"))/100)))+1)));;;"T"));1-ROW(INDIRECT(ADRESS(MATCH(B1;T!$A:$A;0);1;;;"T")));0)⌋
↪  ))
```

```
V!$B$1=200,000
```

```
V!$B$2=SQRT(INT(NORM.INV(RAND();0;25))^2)+1
```

```
V!$B$6=log($B$1;LIGNES(T!$A:$A)-NB.IF(T!$A:$A;""))
```

```
B1= [preceding segment]
```

```
T!= [table of proximity]
```