IXth Congress of the International Society for Dialectology and Geolinguistics
23-27 July 2018, Vilnius, Lithuania

Measures of frequency and dispersion in individual texts

Applied to the morphology of future of Spanish verbs in space and time

Hiroto Ueda, University of Tokyo

1. Introduction

When dealing with the frequency of words in dialectal texts, two statistical measures are used: mean and standard deviation. Both are calculated between several dialectal texts to measure the magnitude of frequency on average and the degree of dispersion observed among several texts treated. In this way we get to obtain a frequency value and another dispersion value between several dialectal texts. For example, among 5 texts, we obtain a mean and a standard deviation calculated between the 5 texts.

On the other hand, we are interested in evaluating the mean and standard deviation of the words observed in a single monolectal text in order to know its magnitude and stability within the same text. Therefore, we acquire frequencies on average and standard deviations of the texts treated. For example, among 5 texts, we obtain 5 means and 5 standard deviation, corresponding to each text.

As for the standard deviation, we will transform it into the normalized standard deviation, which has the range from 0 to 1, which is convenient for comparing the vocabulary of different dimension, for example, articles, prepositions and verbs.

We will explain the method of calculating the individual frequency and dispersion and we will apply it to the Spanish texts of the Middle and Modern Ages in the two main regions: Castile and Aragon. Our theme is the irregular forms of the future of verbs: *poner* 'to put', *tener* 'to have' and *venir* 'to come'.

We will use the data of the CODEA project («*Corpus de documentos Españoles Anteriores a 1800*» ('Corpus of Spanish documents before 1800')[1] and our linguistic analysis system LYNEAL («*Letras y Números en Análisis Lingüísticos*» ('Letters and Numbers in Linguistic Analysis')[2]. We try to analyze the frequencies of occurrence of

[1] Cf.: http://corpuscodea.es/

[2] Cf.:
https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/
http://shimoda.lllf.uam.es/ueda/lyneal/

the linguistic forms found in the historical and dialectal documents. Unlike the data obtained in sociolinguistic or psycholinguistic studies, in which researchers work with properly pre-established and stratified parameters, the historians of the language are forced to use the materials found in the old documents that have been preserved as over time and across linguistic geography, without the parameters distributed *a priori*.

2. Frequency and dispersion

2. 1. Probabilistic frequency

In this section we will discuss the problems that have the different frequencies usually used in corpus linguistic investigations: the Absolute Frequency (AF), the Relative Frequency (RF) and the Normalized Frequency (NF). To save them, we propose to use a new frequency, which we call "Probabilistic Frequency" (PF), based on the theory of statistical probability.

As an example of the Absolute Frequency (AF), we use the search result of the separate and joined forms of Spanish <de> + singular defined article: <de el> in comparison with <de la>, <del> y <dela>, which LYNEAL system has presented with the CODEA data:

AF	1200	1300	1400	1500	1600	1700
de el	**2**	**1**	**0**	**14**	**53**	**23**
de la	89	38	89	242	420	117
del	800	732	1187	1752	1195	389
dela	402	412	680	1088	364	181
Total	1 293	1 183	1 956	3 096	2 032	710

In this table, we notice the big difference that exists between the total frequencies in the final row. Then it is a mistake to easily compare the figures. For example, in the table, we note certain frequencies of <de el> in 1500, 1600 and 1700, whose top is at 1600 (53), followed by 1700 (23). We know that this observation is wrong. We should relativize these concrete figures by the total amount of each column in form of ratio or percentage, which are Relative Frequencies (RF):

RF (%)	1200	1300	1400	1500	1600	1700
de el	**0.2**	**0.1**	**0**	**0.5**	**2.6**	**3.2**
de la	6.9	3.2	4.6	7.8	20.7	16.5
del	61.9	61.9	60.7	56.6	58.8	54.8
dela	31.1	34.8	34.8	35.1	17.9	25.5

In this way, we could confirm the gradual rise of * <de el>. The problem we see in the Relative Frequency is that the proportions are compared equally without seeing the difference of the whole. We wonder if it is correct to compare the figure of 3.2% within 710 and that of 2.6% within 2 032. Said in another simpler way, the question is whether 3 among 12 (25%) is comparable with 25 among 120 (21%) to affirm that the first case (25%) is more significant than the second. Actually by saying that 3 out of 12 represents 25%, we are performing an operation called 'extrapolation' in modern statistics, in the sense that it is assumed by excess that what happens 3 times in 12 trials would occur at least 25 times in 100 trials, applying the same probability to the greatest number. Without needing to resort to the statistical science that warns its error, by our experience of the life, we know that the few experiments do not guarantee the same significance as many tests.

Another problem that the Relative Frequency presents is that all the observation is limited to the interiority of the table. We are comparing internal relative values, without seeing the whole text. This problem would be solved with the Normalized Frequency (NF), made in the following table. First of all we offer the count of the words in each year parameter:

Year	1200	1300	1400	1500	1600	1700
Palabra	73 316	90 987	119 607	185 010	106 694	49 150

To obtain the Normalized Frequency (NF), the system proceeds to the next operation. We limit ourselves to exposing only the first case of * <de el> in 1200:

NF (*<de el> en 1200) = 2 / 73316 * 100 000 = 2.72

The Normalized Frequency has the merit of representing the relative magnitude within the whole text. However, the same problem of incomparability due to the difference of the total numbers of the word still exists, as in the Relative Frequency (RF). Is it justifiable to compare the figure of 53 within 106 694 (year 1600) and 23 within 49 150 (year 1700)? Here we also realize that extrapolation is carried out, which warns modern statistics.

It is ironic to witness the problem of incomparability, due to the difference of the denominating part of the fraction. Precisely because of the existence of the difference of the same part, the two relativized frequencies have been devised, internal (Relative Frequency) and external (Normalized Frequency). If there were no such difference in the base, there would be no need to relativize the frequency values, since the Absolute Frequency itself is, from the beginning, comparable.

When questioning the validity of the two relativized frequencies, we have used the word 'probability', which we consider important when evaluating numerical data. In order to formulate the frequency free of the problem of the base, we will use the binomial probability, which consists in seeing the probability of success or failure. First we consider the Significance (*s*, function S) that gives the *x* successes in *n* trials of an event endowed with the expectative binomial probability (*e*). To simplify the operation, we use Excel function BINOMDIST, which returns the binomial probability with four parameters, *x, n, e* and 1, the last of which, 1, represents the accumulation of probabilities up to the indicated point, in place of the individual probability (with parameter 0):[3]

$$s = \text{BinS}(x, n, e) = \text{BINOMDIST}(x\text{-}1, n, e, 1)$$

In this way we have four interdependent values, *s, x, n, e;* that is, each value is derivable from the other three. To derive the significance *s*, we use the number of successes (*x*), number of trials (*n*) and expectative probability of the event (*e*). In statistical texts, they usually explain events with examples of coin, dice or card. The probability of event of <face> or <cross> of a coin is 0.5, while each figure of a die presents the probability of 1/6 (= 0.167) and the card, 1/13 (= 0.077).

Unlike the coin, the die or the card, the expectative probability (*e*) of the occurrence of words is unknown. Our objective is to approach it from the frequency of success, that is, from the occurrence in the form of the Absolute Frequency (AF) and the total number of the test, which represents the sum of the column and with a certain significance, for example, 99% (0.99).[4] Now we look for the probability of occurrence of words (*e*), from the number of occurrence (*x* = AF), the total number of occurrences (*n*, column total in RF or total of words in the text, NF) and the significance (*s*), which

[3] See Program-1 in Appendix.

[4] The significance can not be 100% (1.000). It is also not advisable to aspire as high as 99.9%, 99.99%, which can be calculated but offer values that are too modest and impractical. After several experiments, we have come to the conclusion that 99% (0.99) is convenient.

we establish in a figure of 99%. The BinE function returns the probability (*e*):[5]

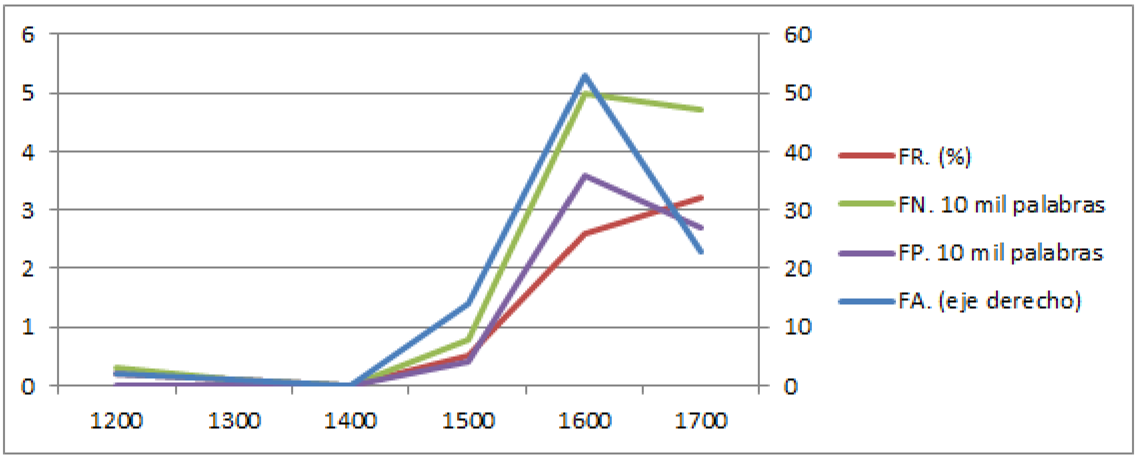
e = BinE(*x, n, s*)

Since the expectative probability (*e*) oscillates between 0 and 1, it is convenient to multiply by a thousand, 10 thousand, 100 thousand, etc., depending on the case, according to the original magnitude of the Absolute Frequency. The original magnitude can be exceeded without fear of falling into the problem of extrapolation, discussed above, since the multiplied number is guaranteed by probability, not dependent on the amount of the denominator, as in the relativized frequencies (RF and NF) .

The same BinE function installed in the LYNEAL system returns the Probabilistic Frequency (PF) in the following way. The multiplier is 10 thousand (words):

PF. 10 mil palabras	1200	1300	1400	1500	1600	1700
de el	0.2	0	0	3.7	35.5	27.1
de la	94	27.7	57.6	112.2	350.5	190.5
del	1004	737.3	927	895.5	1046.5	701.8
dela	487.1	402.8	519.2	547.6	301.1	308.2

Let'*s* compare the four frequencies, AF, RF, NF and PF, of <de el> in the following graph:



In the graph we find outstanding differences in the year 1600 and 1700. In the last year 1700, even the order changes enormously. Naturally the Absolute Frequency (AF) is not comparable. The Relative Frequency (RF) is problematic, since it is limited

[5] See Program-2 in Appendix.

to observing the proportion within the table. The Normalized Frequency (NF) gives some results without seeing the significance, treating all the frequencies in the same way. The Probabilistic Frequency (PF) guarantees the significance of 99% in each figure. The two relativized frequencies, RF and NF, cause problems when the base of comparison, total of the column or total of words, presents great differences and above all, some members of the base are quite small figures, such as one or two digits. The Probabilistic Frequency (PF) is derived from the universal exterior scale of expectative probability that always guarantees the same degree of significance of 99%, both with the reduced base and with the high base. In this sense, the Probabilistic Frequency is robust[6].

2. 2. Block dispersion

In analysis of linguistic data, in addition to observing the existence (or absence) and the frequency of linguistic forms, it is convenient to find out their degree of dispersion with the «StandardDeviation» (SD):

$$SD = \{[(X_1 - M)^2 + (X_2 - M)^2 + \dots + (X_n - M)^2] / N\}^{1/2}$$

where M is mean and N is the data number.

The Standard Deviation, which is used as an indicator of variation, has the property of increasing according to the scale of the data. For this reason, a constant indicator of independent variation of the data scale has been sought. Therefore, the Coefficient of Variation (CV) is calculated from the Standard Deviation (SD) divided by the Mean (M).

$$CV = SD / M$$

Since the Coefficient of Variation (CV) is not normalized, that is, it does not have range between 0 and 1, we have looked for a normalized indicator of variation, which we call «Normalized Standard Deviation» (NSD), which is calculated by the division of the Standard Deviation (SD) by the maximum value of the same Standard Deviation (SD.max):

$$NSD = SD / SD.max$$

Let'*s* see how to find the formula of SD.max. We start with the formula of the standard deviation (SD):

[6] For details see Appendix, which is English translation of the part of Ueda (2017).

$$SD = \{[(X_1 – M)^2 + (X_2 – M)^2 + \ldots + (X_n – M)^2] / N\}^{1/2}$$

Suppose we are dealing with a set of data with an extreme case of deviation, eg. (10, 0, 0, 0, 0), which logically has the maximum value of Standard Deviation (SD.max) To generalize the problem, we use K instead of a specific number (10): (K, 0, 0, ..., 0). Then, only the first term of SD is $(K - M)^2$, and all the remaining, N - 1 cases, are $(0 - M)^2 = M^2$, and therefore, the maximum value of Standard deviation (SD. max.) is:

$$SD.max. = \{([(K - M)^2 + (N - 1)\ M^2]\}^{1/2}$$

where, K is equal to the sum of the data, since the rest are null. Since the sum is equal to the Mean (M) multiplied by the Number (N) of data (Sum = N M ← M = Sum / N), K is equal to N M:

$$K = Sum = N\ M$$

Thus:

$$\begin{aligned}
SD.max &= \{[(N\ M – M)^2 + M^2 (N – 1)] / N\}^{1/2} && \leftarrow\ K = N\ M \\
&= \{[(M\ (N – 1))^2 + M^2 (N – 1)] / N\}^{1/2} && \leftarrow\ M \text{ al exterior} \\
&= \{[\underline{M^2 (N – 1)^2} + \underline{M^2 (N – 1)}] / N\}^{1/2} && \leftarrow\ M^2 \text{ es común} \\
&= \{M^2 (N – 1)\ [\underline{(N – 1) +1}] / N\}^{1/2} && \leftarrow\ M^2 (N – 1) \text{ es común} \\
&= \{(M^2 (N – 1)\ \underline{N / N}\}^{1/2} && \leftarrow\ (N – 1) +1 = N \qquad \leftarrow (N-1) +1=N \\
&= [(M^2 (N – 1)]^{\underline{1/2}} && \leftarrow\ N / N = 1 \\
&= M\ (N – 1)^{1/2} && \leftarrow\ (M^2)^{1/2} = M
\end{aligned}$$

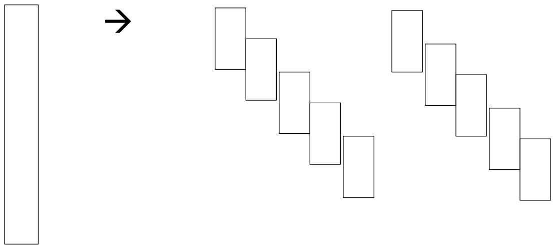
Therefore, the Normalized Standard Deviation (NSD) is:

$$NSD = SD / SD.max = SD / [M\ (N - 1)^{1/2}]$$

Now, instead of looking for the deviation between variables in the form of several texts (A), we can calculate the degree of dispersion by dividing the text in, for example, 10 blocks (B).

A: □ □ □ □ □ □ □ □ □ □

B:

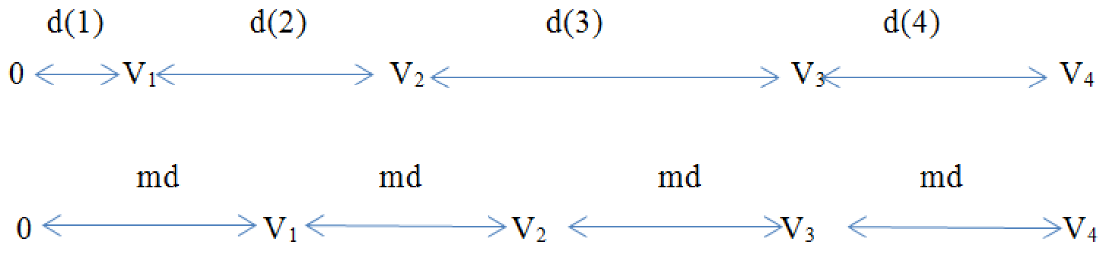


The complement of the «Standard Normalized Deviation in Blocks» (NSD.b) is Block Dispersion (Disp.b):

Disp.b. = 1 - NSD.b

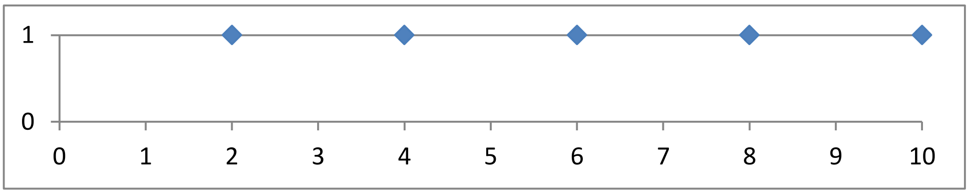
2. 3. Uniformity

To study the degree of dispersion within a material, we propose a formula that we call "Coagulation" and its complement "Uniformity", which consists of comparing the distances between the elements, d (1), d (2), ..., with the mean distance (md) that is calculated by dividing the total length by N (number of elements):



First, let's see a distribution completely matched, for example, the word *ah* in "*Yesterday ah we ah saw ah my ah friend ah*". In total we have 5 occurrences of *ah* in positions (2, 4, 6, 8, 10). All distances are equal and coincide with the average distance (10 / 5 = 2) and all differences between the actual distance and the average distance are zero (0). The «Sum of the Differences between real distance and average distance» (S.Dist.) is zero (0). In this case the Coagulation is null (0) and the Uniformity is complete (1):

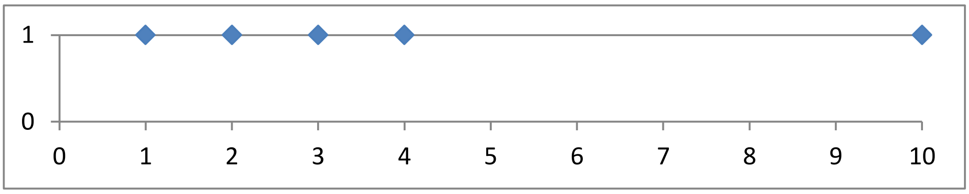
Element	Position	Distance	Mean distance	|Dis. -M. Dis.|
x1	2	2	2	0
x2	4	2	2	0
x3	6	2	2	0
x4	8	2	2	0
x5: M	10	2	2	0



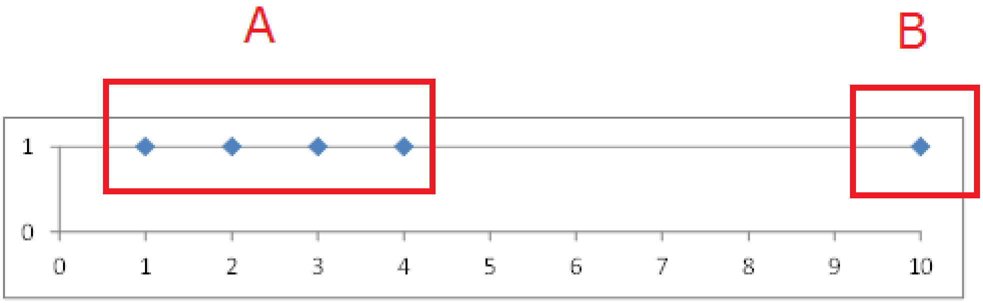
Position (2, 4, 6, 8, 10), Coagulation: 0.000, Uniformity: 1,000

The maximum Coagulation is obtained when all the elements are joined except the final one (x5), which distances itself from the remains (x1, x2, x3, x4):

Elemento	Posición	Distancia	Dist.media	|Dis. - Dis.media
x1	1	1	2	1
x2	2	1	2	1
x3	3	1	2	1
x4	4	1	2	1
x5: M	10	6	2	4



Position (1, 2, 3, 4, 10), Coagulation: 1,000, Uniformity: 0.000



To calculate the Sum of the Difference between the real distance and the average distance (S.Dist), we divide the points into two parts: the first part (A) with x1, x2, x3, x4 and the second (B) with x5. In the first part (A), we have 4 elements and the maximum of the sum of the distances (S.Dist.max) is::

SD.max (A) = (5 - 1) * | 1 - 10 / 5 | = 4 * | 1 - 2 | = 4

where 4 times (5-1) of the difference between the real distance (1) and the average distance (2) are calculated. To generalize the calculation, we use the letter N for the number of elements (5) and the letter M, the maximum amount of all the elements (10):

$$\begin{aligned}\text{S.Dist.max(A)} &= (N - 1) * |1 - M / N| \\ &= (N - 1) * |N - M| / N \\ &= (N - 1) * (M - N) / N \quad \leftarrow M \geq N\end{aligned}$$

For the second part (B), the maximum of the Sum of the differences (S.Dist.max) is:

$$\text{S.Dist.max (B)} = |\ 10 - (5 - 1) - 10/5\ | = |\ (10 - 4) - 2\ | = 4$$

where (5 - 1) indicates point 4 of the graph and the difference between the actual distance (10 - 4) and the average distance (10/5 = 2) is 4.
Curiously, the S.Dist.max (A) and S.Dist.max (B) coincide, being both 4. We find it out with the generalized formula:

$$\begin{aligned}\text{S.Dist.max(B)} &= |M - (N - 1) - M / N| \\ &= |N * (M - N + 1) - M| / N \\ &= |N * M - N * N + N - \underline{M}| / N \\ &= |N * M - \underline{M} - N * N + N| / N \\ &= |M * (N- 1) - N * (N - 1)| / N \\ &= |(N - 1) * (M - N)| / N \\ &= (N - 1) * (M - N) / N \quad \leftarrow N \geq 1, M \geq N\end{aligned}$$

In this way, the theoretical maximum of the sum of the distances (S.Dist.max) is:

$$\begin{aligned}\text{S.Dist.max} &= \text{S.Dist.max (A)} + \text{S.Dist.max (B)} \\ &= 2 * (N - 1) * (M - N) / N = 4 * 2 = 8\end{aligned}$$

We define the degree of Coagulation:

$$\text{Coagulation} = \text{S.Dist} / \text{S.Dist.max} = \text{S.Dist} / [2 * (N-1) * (M-N) / N]$$

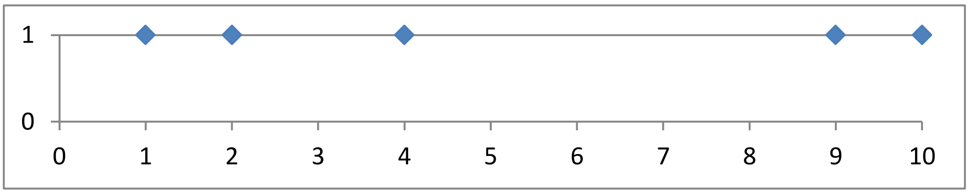
The complement of Coagulation is «Uniformity»:

$$\text{Uniformity} = 1 - \text{Coagulation} = 1 - \text{S.Dist} / \text{S.Dist.max}$$

As an example, we calculate Coagulation and Uniformity of a data set (1, 2, 4, 9, 10):

Element	Position	Distance	Dist.mean	|Dis. - Dis.mean|
x1	1	1	2	1
x2	2	1	2	1
x3	4	2	2	0
x4	9	5	2	3
x5: M	10	1	2	1

The sum of the distance is 1 + 1 + 0 + 3 + 1 = 6. The Coagulation is 6 / 8 = .750 and the Uniformity is 1 - .750 = .250:



Position (1, 2, 4, 9, 10), Coagulation: .750, Uniformity: .250

2. 4. Stable use

Juilland and Chang Rodríguez (1964) proposed the formula of Usage (U) of words by multiplying the frequency (F) of the word in question by its degree of Dispersion (Disp), with the five data sets: dramas, novels , essays, scientific documents and news:

U = F * Disp

Considering the degree of use of words, the two authors treat not only their frequencies but also the degrees of dispersion and presented the following formula of the degree of dispersion (Disp):

Disp = 1 - SD / (2 * Mean)

We find that the number 2 in the denominator represents: (the number of variables 5 - 1)$^{1/2}$. In this way, we note that SD / (2 * Mean) is the Normalized Standard Deviation (NSD). Therefore, to generalize the value of Dispersion (Disp), we will use the following formula;

Disp = 1 - NSD

We think that the same method is applicable for the analysis of a text or set of texts, without variables. For our part, we propose to include not only the Dispersion, but also the Uniformity for the calculation of Stable Usage (SU), which we define as follows:

$$SU = F * (Disp.b * Unif)^{1/2}$$

The reason why we take into consideration the two figures, Block Dispersion (Disp.b) and Uniformity (Unif), is based on the fact that the two indices complement each other in order to arrive at an unbiased interpretation. The Block Dispersion serves to see the distribution of frequencies between the blocks as a whole, while the Uniformity serves to see the vicissitudes of individual distances.

3. Application

3. 1. Epenthetic forms

As an example of application, we propose to analyze the forms of future and conditional verbal conjugation in medieval and modern Spanish. The manuals of Spanish historical grammar usually list the future variants of the verbs, *tener* 'to have', *poner* 'to put', *venir* 'to come': *tenrá, terné, tendré, tendría, venr(r)án, vernié* etc. Lloyd (1987: 496-7) and Penny (2006: 242) classify them with terms of "reinforcement" (*venr(r)án* 'they will come'), "metathesis" (*terné* 'I will have', *verníe* 'I would come'), "epenthesis" (*tendrá* 'he will have') and "assimilation" (*porrá* 'he will put', *verrán* 'they will come').

We are interested not in a simple list but in its distribution in space and time. We are now in a condition to approach it by means of the diachronic corpus that covers a wide geography of the Spanish language and a long chronology of centuries. We refer to the CODEA corpus that we have presented in the introduction of this work. To take advantage of this, we have installed frequency and dispersion calculation functions with the graphics in our LYNEAL system, which we also have presented in the introduction.

Especially we are interested in the current forms *pondré* 'I will put', *tendre/* 'I will have', *vendré* 'I will come', with the epenthesis of D between N and R in the form of NDR. The geographical distribution of its first appearance, according to CODEA data, shows the early supremacy in the eastern region of the Iberian Peninsula, specifically in the Kingdom of Aragon, while its appearance in the central and western regions, in the territory of the Corona de Castile is quite late.

Moreno Bernal (2004: 155), who has researched his corpus of literary texts, suggests the "Aragonese influence":

> The Aragonese texts adopt the epenthetic solution for the futures of *poner, tener, venir* before the Castilians. The forms *pondré, tendré, vendré* 'I will put, I will have, I will come' that appear in Castilian texts prior to the fifteenth century are usually the product of the Aragonese influence (so *pondrán* 'they will put' and *avendremos* 'we will come' in Cid).

Unlike literary texts, in notarial documents, the forms of NDR in Castile are scarce as we will see in the subsequent sections. Marina Serrano confirms this fact in her recent thesis (2018) with the texts of «CODEA». Effectively in the kingdom of Aragon, its first appearance in the same corpus is in Teruel (1277), followed by Huesca (1285) and Zaragoza (1406). Barcelona in 1481 also precedes the Castilian provinces, where they begin to use the epenthetic forms in later centuries, mostly in the sixteenth and seventeenth centuries, as described in the following map[7]:

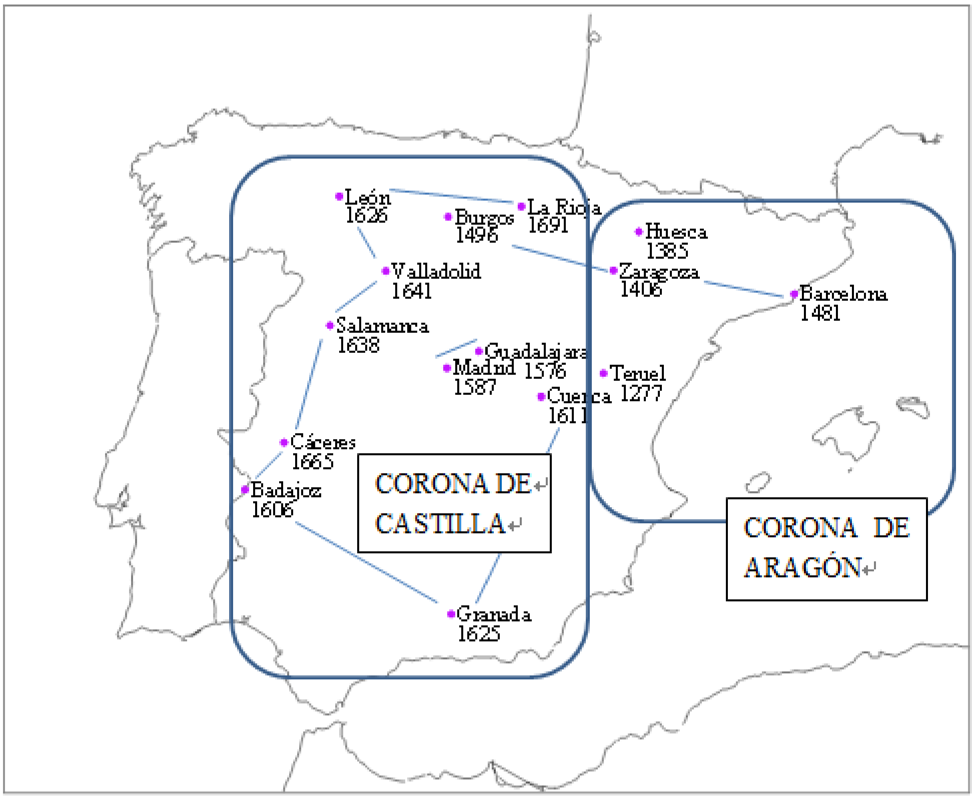


Fig. 3.1. First appearance of the NDR form

[7] For the division between the territories of Castile and Aragon, we have consulted López-Davalillo Larrea (2000: 123, 135).

3. 2. Aragon and Navarra

Based on the information obtained in the previous section, it is convenient to go through history not mixing the two well differentiated regions: Castile and Aragon. In the first place we will see the chronology of the variants in Aragon, which precede Castile in the first appearance of NDR,. The kingdom of Navarre, situated between Castile and Aragon, we include in the group of Aragon, for possessing more dialectal affinity with this region than Castile.

The CODEA corpus collects different numbers of Aragonese documents distributed over the centuries. Each chronological strip with interval of 50 years is indicated with the year of beginning, for example, 1250 comprises from 1250 to 1299:

Year	1250	1300	1350	1400	1450
Document	8	16	39	39	14
Word	4 560	8 853	28 591	41 346	19 196

Table 4.2.a. Number of documents and words in Aragon and Navarra

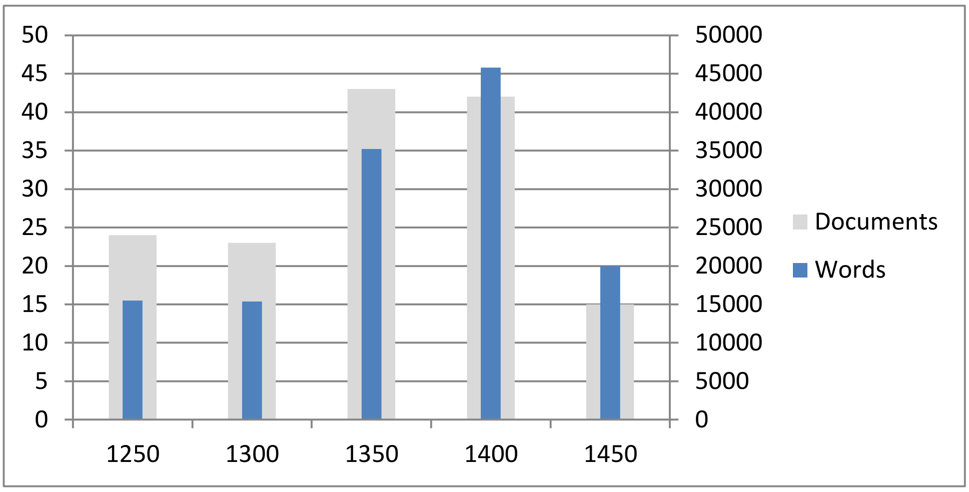


Fig. 4.2.a. Number of documents and words in Aragon and Navarra

The following table represents the absolute frequency of each corresponding form:

Year	1250	1300	1350	1400	1450
NR	3	0	23	4	0
RN	0	3	7	32	17
NDR	1	0	6	19	17
RR	6	2	7	1	1

Table 4.2.b. Future forms in NR, RN, NDR, RR in Aragon.
Absolute frecuency

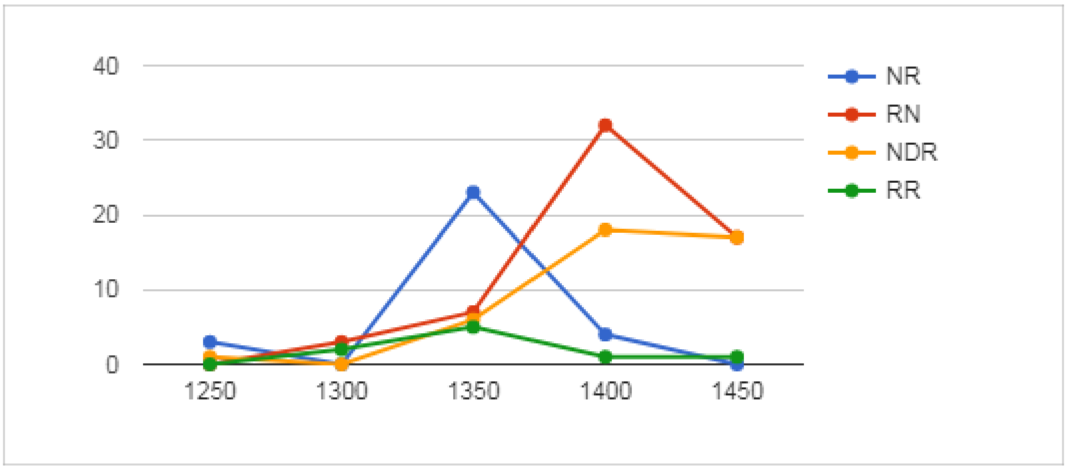


Fig. 4.2.c. Future forms in NR, RN, NDR, RR in Aragon.
Absolute frecuency

Absolute frequencies are not comparable because of their unequal dimension of words in total. Therefore, we use the probabilistic frequency:

Year	1250	1300	1350	1400	1450
NR	2.81	.00	37.91	1.76	.00
RN	.00	2.81	6.63	44.39	44.58
NDR	.05	.00	5.10	22.55	44.58
RR	11.49	1.00	6.63	.05	.05

Fig. 4.2.c. Future forms in NR, RN, NDR, RR in Aragon
Probabilistic frequency for 100 thousand words

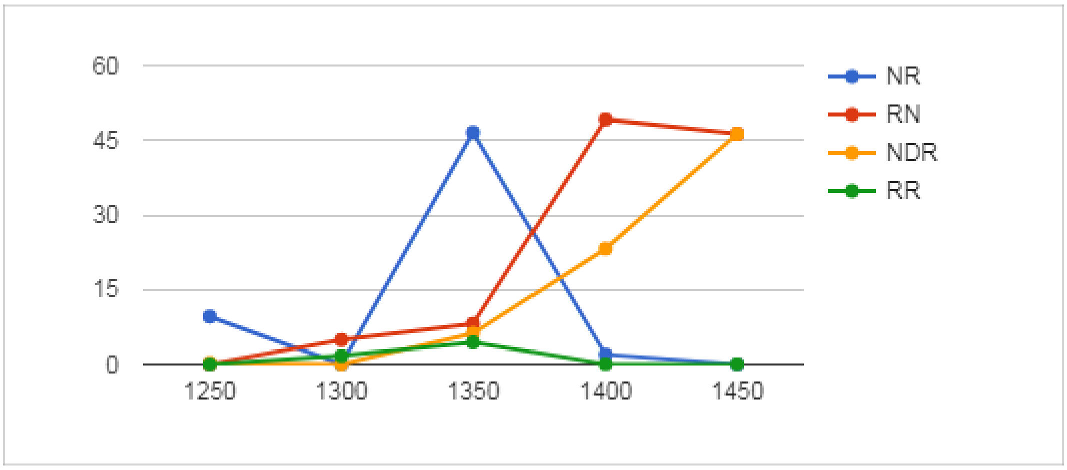


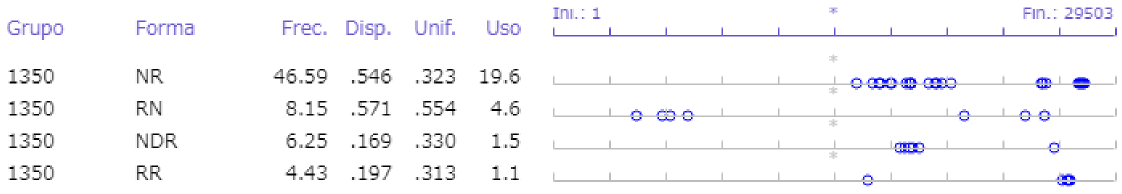
Fig. 4.2.d. Future forms in NR, RN, NDR, RR in Aragon

Probabilistic frequency for 100 thousand words

According to the probabilistic frequency, we can establish the numerical supremacy of NR in 1350, followed by RN and NDR. The assimilated form RR is minority. As a working hypothesis we can think that both the RN metathesis form and the NDR epenthesis form are born from the same origin of NR. Also the RR minority form we assume from NR by regressive assimilation rather than from RN by progressive assimilation. It is due to two reasons: regressive assimilation is more common than progressive assimilation in general, and also the RR chronology does not coincide with that of RN, but rather with that of NR.

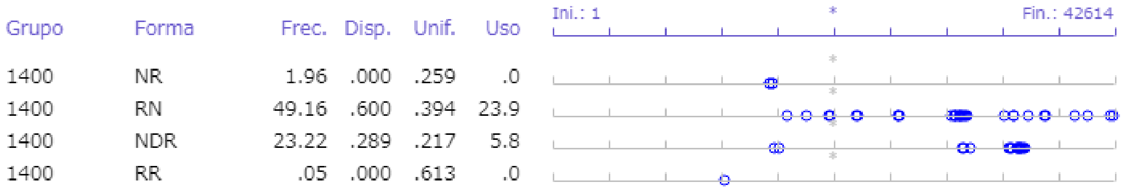
$$\begin{tikzcd} & \text{RN} \\ \text{NR} \arrow[ur] \arrow[r] \arrow[dr] & \text{NDR} \\ & \text{RR} \end{tikzcd}$$

Regarding the dispersion, let'*s* see it in the three important stages: 1350, 1400 and 1450:



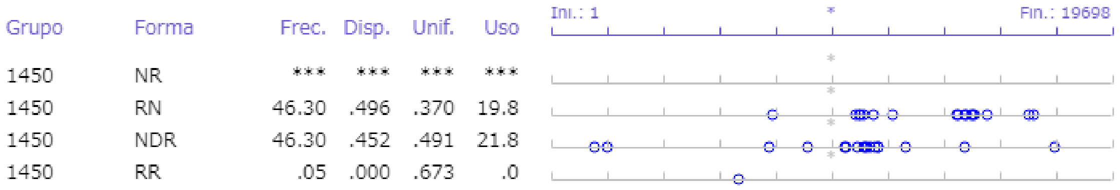
Grupo	Forma	Frec.	Disp.	Unif.	Uso
1350	NR	46.59	.546	.323	19.6
1350	RN	8.15	.571	.554	4.6
1350	NDR	6.25	.169	.330	1.5
1350	RR	4.43	.197	.313	1.1

Fig. 4.2.*e*. Frequency and dispersion of RN, NDR, RR in Castile of 1350



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1400	NR	1.96	.000	.259	.0
1400	RN	49.16	.600	.394	23.9
1400	NDR	23.22	.289	.217	5.8
1400	RR	.05	.000	.613	.0

Fig. 4.2.f. Frequency and uniformity of RN, NDR, RR in Castile of 1400



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1450	NR	***	***	***	***
1450	RN	46.30	.496	.370	19.8
1450	NDR	46.30	.452	.491	21.8
1450	RR	.05	.000	.673	.0

Fig. 4.2.*g*. Frequency and uniformity of RN, NDR, RR in Castile of 1400

In 1350, the dispersion of the NDR epenthetic form is almost nil because it appears in a single block, which is improved in subsequent stages. In 1450 the same form becomes quite stable, being more so than the form of metathesis RN. The epenthetic form was born (1250) and prospered (1400) in the Aragonese region.

3. 3. Castile

In this section we will observe the vicissitudes of the same forms in the documents issued in the Kingdom of Castile. The following table shows the numbers of words and documents collected in the CODEA corpus:

Year	1200	1250	1350	1400	1450	1500	1550	1600	1650	1700	1750
Doc	43	176	81	81	101	158	348	184	131	122	193
Word	18976	98139	87449	97972	112015	128899	145371	78822	94090	53259	72603

Table. 4.3.a. Number of documents and words in Castile

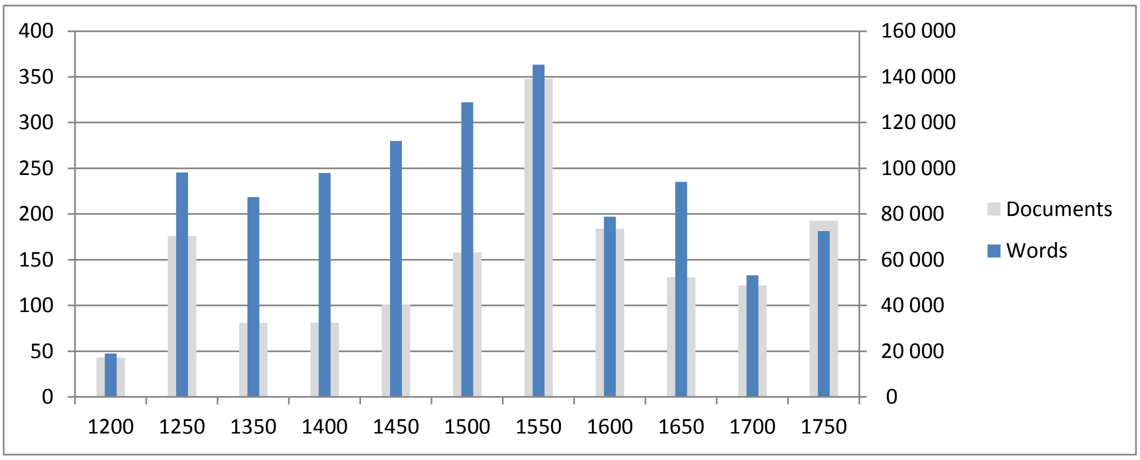


Fig. 4.3.a. Number of documents and words in Castile

We emphasize the fact that in Castile there has not been any case of NR, as shown by the table of absolute frequency:

Year	1200	1250	1350	1400	1450	1500	1550	1600	1650	1700	1750
RN	10	9	7	13	28	19	33	10	1	0	0
NDR	0	0	0	0	2	0	20	13	13	21	29
RR	0	1	0	0	0	0	0	0	0	0	0

Fig. 4.3.b. Future forms in RN, NDR, RR in Castile

Absolute frecuency

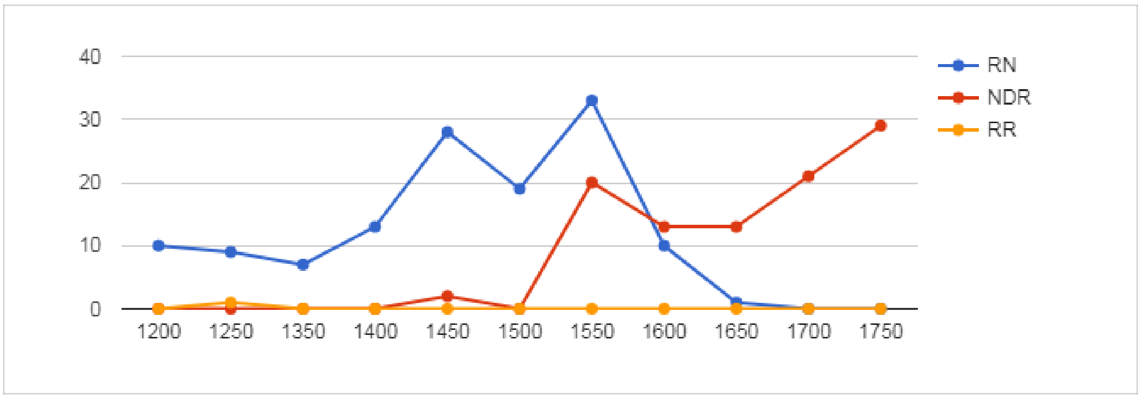


Fig. 4.3.b. Future forms in RN, NDR, RR in Castile
Absolute frecuency

We confirm with documentary evidence that in Castile from the beginning the form of RN metathesis prevails and later in 1600 it yields to the new epenthetic form NDR. The minority form RR disappears completely in the Middle Ages (1350). The same is confirmed in the probabilistic frequency:

Year	1200	1250	1350	1400	1450	1500	1550	1600	1650	1700	1750
RN	21.79	3.58	2.62	6.25	15.31	8.06	14.54	5.20	.05	.00	.00
NDR	.00	.00	.00	.00	.14	.00	7.58	7.77	6.44	22.17	24.75
RR	.00	.05	.00	.00	.00	.00	.00	.00	.00	.00	.00

Fig. 4.3.c. Future forms of conjugation in RN, NDR, RR in Castile
Probabilistic frequency per 100,000 words

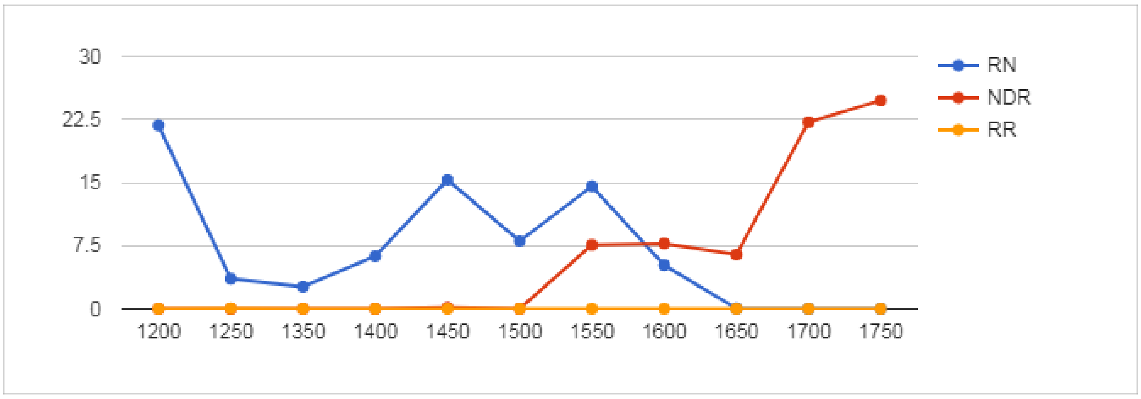
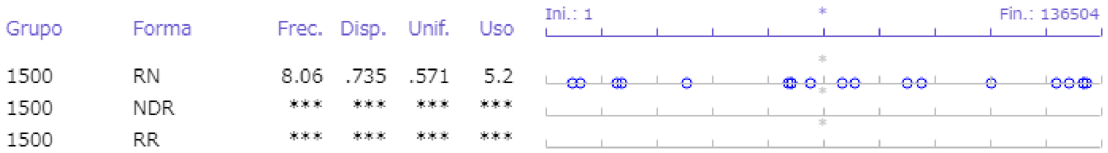


Fig. 4.3.c. Future forms of conjugation in RN, NDR, RR in Castile
Probabilistic frequency per 100,000 words

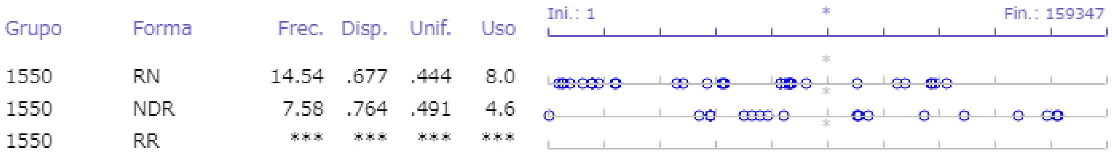
Let's see the probabilistic frequency along with its Dispersion in block and Uniformity. In 1500, only the form of RN metathesis is observed quite stable:



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1500	RN	8.06	.735	.571	5.2
1500	NDR	***	***	***	***
1500	RR	***	***	***	***

Fig. 4.3.d. Frequency and dispersion of RN, NDR, RR in Castile de 1500

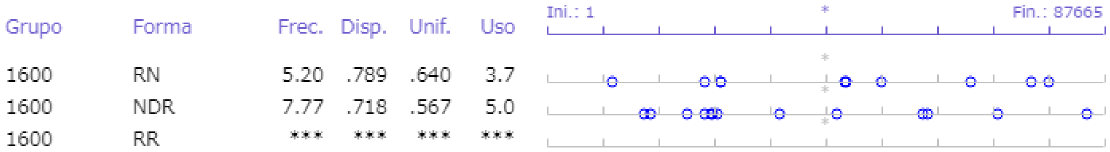
In 1550 appears the new epenthetic form NDR, quite stable, in competition with the old RN:



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1550	RN	14.54	.677	.444	8.0
1550	NDR	7.58	.764	.491	4.6
1550	RR	***	***	***	***

Fig. 4.3.*e*. Frequency and dispersion of RN, NDR, RR in Castile of 1550

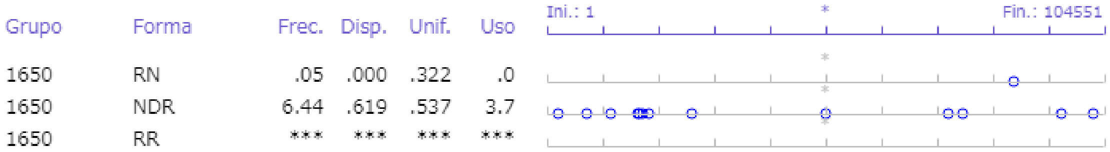
In 1600 the old form RN decays:



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1600	RN	5.20	.789	.640	3.7
1600	NDR	7.77	.718	.567	5.0
1600	RR	***	***	***	***

Fig. 4.3.f. Frequency and dispersion of RN, NDR, RR in Castile of 1600

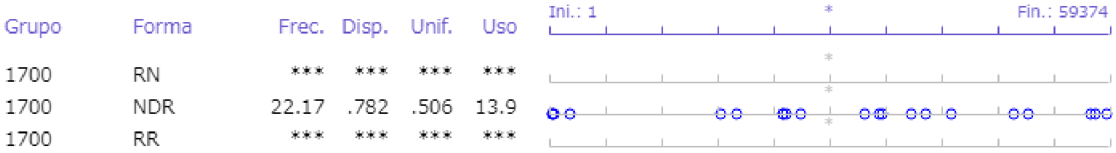
In 1650 the old form RN is almost non-existent in favor of the new NDR:



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1650	RN	.05	.000	.322	.0
1650	NDR	6.44	.619	.537	3.7
1650	RR	***	***	***	***

Fig. 4.3.*g*. Frequency and dispersion of RN, NDR, RR in Castile of 1650

Finally in 1700, the old form RN disappears and the new form NDR maintains its stability:



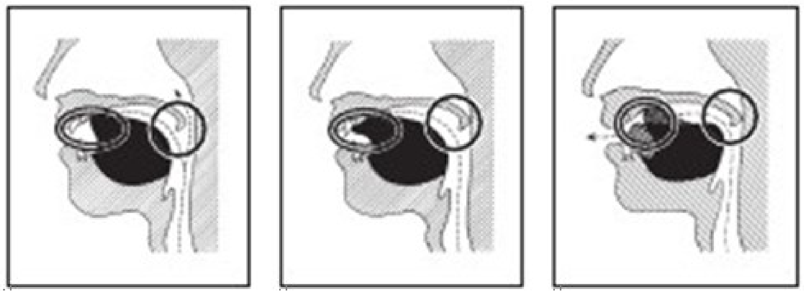
Grupo	Forma	Frec.	Disp.	Unif.	Uso
1700	RN	***	***	***	***
1700	NDR	22.17	.782	.506	13.9
1700	RR	***	***	***	***

Fig. 4.3.h. Frequency and dispersion of RN, NDR, RR in Castile of 1650

4. Discussion

We have observed the change of the verbal form of the future of the old RN in the new NDR, for example, from *terné* to *tendré*, in Castile of 1600. The appearance of the new epenthetic form dates back to 1550. We think that it is very difficult to explain it as a phonological change: /r *n*/ > /*n* d r/. Nor is it possible to resort to the common origin, / *n* r /, since in Castile the same forms, as *ponré, tenré, venré* ('I will put', 'I will have', 'I will come'), are not observed in any document.

The Aragonese situation is different, where the syncopated form with NR was frequent, especially in 1350: *tenrá, tenrán, venrán* ('he will have', 'they will have', 'they will come'), in Huesca from 1275 to 1379. From this syncopated form, it is easy to derive the epenthetic forms, *tendrá, tendrán, vendrán,*since in the articulation of [*n*] buccal occlusion is performed in the dentoalveolar area and the nasal opening and in that of [r] the buccal opening in the form of apical vibration and the closure of the nasal route. Between the two operations can intervene an intermediate step with the nasal and buccal closure at the same time, which produces the articulation of the consonant [d].



[*n*] [d] [r]

In Medieval Castilian, practically the only existing form was that of metathesis, RN, from which it is impossible to directly derive the new epenthetic form with NDR. We think that this form was probably introduced from the Aragonese region. It is about lexical transfer rather than phonological change.

It is well known that within Romanesque languages of Spain, Galician retains inner vowels and Catalan loses them. Within the general scope of Romance languages, Portuguese preserve it and Italian, French, Provençal lose it. Spanish, in its Castilian variant, vacillates between the two extremes (Ueda 2015).

Among Romance languages in which the inner vowel has been deleted, Lausberg (1973: 379) compares the forms from the Latin VENIRE HABEO 'I will come': it. *verrò*, fr. a. *vendrai*, fr. m. *viendrai*, prov. a. *venrai*, cat. *vindré,* esp. a. *verné* y esp. m. *vendré*, port. *virei.*[8] By placing these forms in the geography of the Romance languages, we can see that the oldest forms with R are found in Galician and Portuguese, where there was no vowel fall and the lack of interior N is due to its peculiar phonological characteristic of these languages. Another old form with NR is found in Provencal, which is at the midpoint between French, Catalan with extension to Spanish with NDR, on the one hand, and Italian with RR, on the other. The form of metathesis in ancient Spanish is typical of the same language:

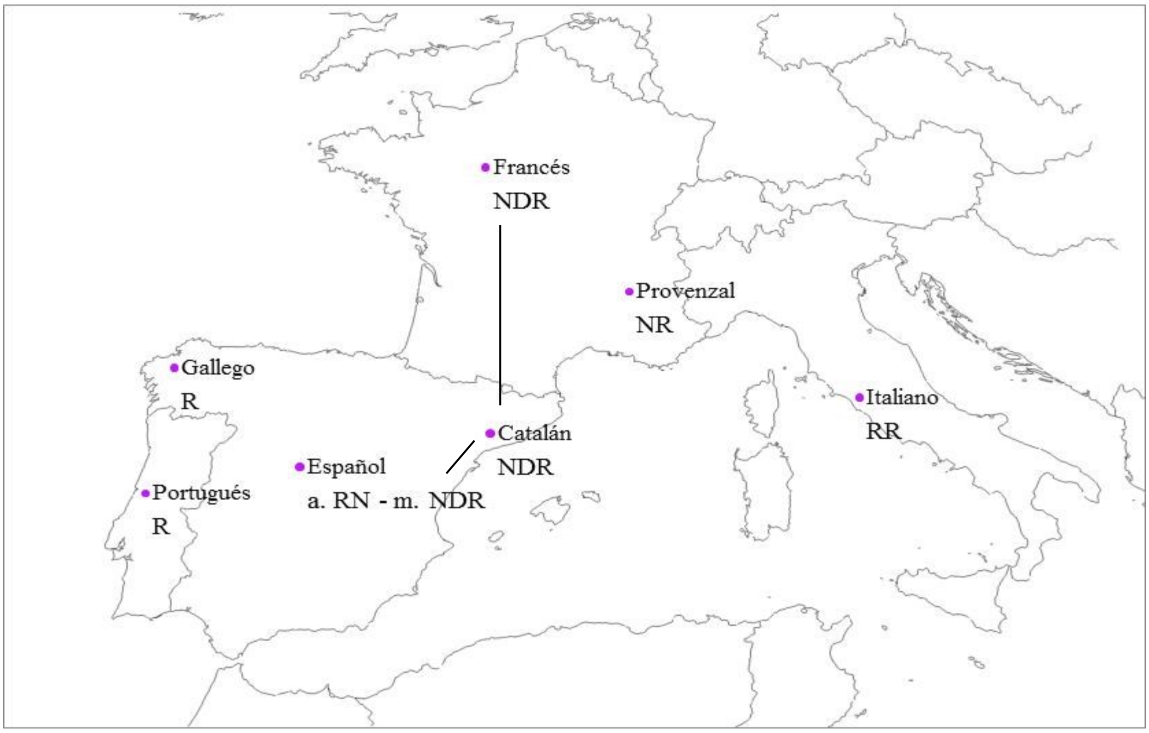
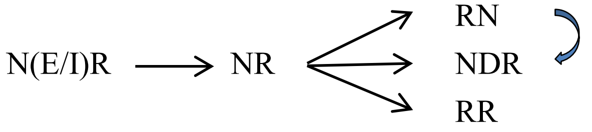


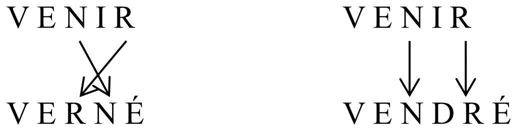
Fig. 5. Future forms with R, RN, NDR, NR, RR in Western Romania

In the Romance languages, the chronological change and the syntactic variation follow the transition from N(E/I)R > NR > RN, NDR, RR. Castillian dialect has introduced the Aragonese form (Moreno Bernal 2004; Serrano 2018), which possessed the same characteristic of Catalan with NDR. Spanish had no problem introducing the new NDR form, since the same sequence was natural and frequent (*Andrés, Alexandro, almendro, honra* > *hondra*). In the following graph the straight arrows indicate phonological changes, while the vertical curve, the lexical transfer:

[8] The retorromance and the Romanian use periphrasis with the verb VENIRE 'to come' > (rh.*vegnir*) and VELLE 'to want' > VOLERE> (ru.a voi), respectively. c. Kataoka (1982: 305-6).



The reason why the Castilians preferred the new form NDR (*vendré* 'I will come') instead of the old one RN (*verné*) is found in the analogy with other forms of the same verbal paradigm. The future with the sequence of R-N was different from the original sequence N-R in the infinitive *venir* 'to come'. The new form with N-D-R satisfies the condition of the order of N-R, although with the additional insertion of D. We think that the insertion of an element, NR> NDR, would not break both the formal regularity and the drastic change of order RN > RN. Alvar and Pottier (1983: 251) indicate: "the normal thing was that the language tried to maintain the lexematic uniformity, broken by all these realizations (*com-er*, but *comb-ré; ven-ir,* but *vend-ré,* etc.)". And we think that *vend-ré* with epenthesis is "more uniform" with the lexeme *venir*, than *verné* with metathesis:



Another reason of preference for epenthesis rather than metathesis is that there are numerous cases of similar epenthesis (Peny 2006: 111): M'N >MBR (HOM(I)NE > *hombre* 'man'; FEM(I)NA > *hembra* 'female'; SEM(I)NARE > *sembrar* 'to seed'), M'R > MBR (HUM(E)RU > *hombro* 'shoulder'), M'L > MBL (TREM(U)LARE > *temblar* 'to shake'), en contraste con los contados casos de metátesis N'R > RN (GEN(E)RU > *yerno* 'son-in-law', VEN(E)RIS > *viernes* 'Friday'). Menéndez Pidal (1972: 160-161) lists more cases in the section "Grupos interiores romances" (Romance interior groups) , while dealing with cases of metathesis in the chapter "Cambios fonéticos esporádicos" (Sporadic phonetic changes) (185), where he states that "la R es la más insegura" (the R is the most insecure). That is, the epenthesis is general, while the metathesis is special.

5. Conclusion

We owe our knowledge of the ancient forms of verbal future to the manuals of Spanish historical grammar. We had not known, however, the chronological vicissitudes and the dialectal variations, since the manuals expose the forms in a juxtaposed way, without offering precise information in space and time. For example, Hanssen'*s* old historical grammar (1913: 119) explains: "Some interesting forms are the following: *terné, porné, verné* (var. *terré, tenré,* etc.) by the side of *tendré*, etc." In our view, the forms in RN and RR were not by the side of the forms in NDR. They were distributed in different manner, both chronologically and geographically. In Castile the order of RN to NDR is fundamental. We have observed the phonological change from NR to NDR in Aragon and the lexical transfer from RN to NDR in Castile. In both regions, it is not about the coexistence of both variants but the substitution of the old form for the new one.

Thanks to the new data presented by the diachronic corpus project CODEA, we can now try to approach the historical-geographical reality. To statistically evaluate the documentary evidence, we have installed the functions of the Probabilistic Frequency, which offer a high degree of significance (99%) and the Block Dispersion and Uniformity, which confirm the degree of stability of the recorded frequencies. We believe that we have demonstrated the usefulness of both the materials and the tool in dealing with the subject that characterizes the Spanish language.

Reference

Alvar, Manuel / Pottier, Bernard. 1983. *Morfología histórica del español.* Madrid. Gredos.

Andrés Díaz, Ramón de. 2013. *Gramática comparada de las lenguas ibéricas.* Gijón, Ediciones Trea.

Hanssen, Federico. 1913, 1966. *Gramática histórica de la lengua castellana*. París. Ediciones Hispano-americanas.

Juilland, Alphonse / Chang-Rodríguez, E. 1964. *Frequency dictionary of Spanish words,* The Hague, Mouton.

Kataoka, Kozaburo. 1982. *Romansugo rekishi bunpou. (Gramática histórica de las lenguas románica.)* Tokio. Asahishuppansha.

Lapeza, Rafael. 1980. *Historia de la lengua española*. Madrid, Gredos.

Lausberg, Heinrich. 1976. *Lingüística románica. Tomo II. Morfología*. Madrid, Gredos.

Lloyd, Paul M. 1987. *Del latín al español. I. Fonología y morfología históricas de la lengua española*, Madrid, Gredos.

López-Davalillo Larrea, Julio. 2000. *Atlas histórico de España y Portugal. Desde el Paleolítico hasta el siglo XX.* Madrid. Editorial Sintesis.

Menéndez Pidal, Ramón. 1968. *Manual de gramática histórica española,* 13ed. Madrid, Espasa-Calpe.

Moreno Bernal, Jesús. 2004. "La morfología de los futuros románicos. Las formas con metátesis". en *Revista de Filología Románica*, núm. 21, pp. 121-169.

Penny, Ralph. 2006. *Gramática histórica del español*. Barcelona, Ariel.

Serrano Marín, Marina. 2018. *Estudio de la morfología verbal del español en fuentes documentales de los siglos XIII-XVI,* Tesis doctoral presentada en la Universidad de Alcalá.

Ueda, Hiroto. 2015. «La vocal débil en la apócope extrema medieval: Observaciones sobre el Corpus de Documentos Españoles Anteriores a 1700», en Sánchez Méndez, J. P., M. de la Torre / V. Codita (eds.) *Temas, problemas y métodos para la edición y el estudio de documentos hispánicos antiguos.* Valencia: Tirant Humanidades, pp. 585 ‑ 607.

_____. 2011. *Supeingo bunpou handobukku. (Manual de gramática española).* Tokio. Kenkyusha.

_____. 2017. *Análisis de datos cuantitativos para estudios lingüísticos.* https://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/4-numeros/doc/numeros-es.pdf

_____ / Moreno Sandoval, Antonio. 2018. «Unión y separación de preposición y artículo definido del español. Observaciones con la frecuencia probabilística en el análisis de Pareto», comunicación oral presentada en el *X Congreso Internacional de Lingüística de Corpus*, Cáceres, España, 9 de mayo, 2018.

Appendix: Probabilistic frequency

Problem

In corpus linguistics, with multiple search (in several forms at the same time) with multiple attributes (for example, years), we get to obtain a two-dimensional table, of forms and variables (years). Now we can see the phenomenon not only in a few years (1200, XIII century), but to observe the linguistic changes along the chronology (1200, 1250, etc.) compared to other forms. Let'*s* see a real data of the Absolute Frequency (AF) of the three forms with orthographic variation: <uoz>, <boz>, <voz>[9]:

[9] The table was obtained on the "CODEA in LYNEAL" site: http://shimoda.lllf.uam.es/ueda/lyneal/codea.htm

AF	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Sm	3	12	12	52	94

These frequencies, however, are not comparable, since the Sum (Sm) of the three forms are different {3, 12, 12, 52, 94}. For example, the figure 3 <uoz> in 1200 is not comparable with 8 in the same way in 1250. In this case, the researchers resort to the Relative Frequency (FR), which is calculated by dividing the Absolute Frequency by the Sum, for example, 3 / 3 = 1,000, 8 / 12 = .667. If we multiply the Relative Frequency by 100, we arrive at the percentage: 1.000 * 100 = 100 (%), 0.667 * 100 = 66.7 (%):

AF	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Sm	3	12	12	52	94

FR (%)	1200	1250	1300	1350	1400
uoz	**100.0**	**66.7**	25.0	21.2	6.4
boz	0.0	25.0	66.7	34.6	37.2
voz	0.0	8.3	8.3	44.2	56.4
Sm	100.0	100.0	100.0	100.0	100.0

However, neither the Relative Frequency (FR) nor the percentage (%) are adequate to compare the figures with different bases. For example, 3 out of 3 (FR: 1.000 (100%)) presents the figure greater than 8 out of 12 (FR: 0.667 (66.7%)), even though we think and intuit that 3 out of 3 is less important than 8 among 12 and much less important than 80 among 120. Football fans know that the player who has scored 3 goals in 3 games is less important than the other who has scored 8 goals in 12 games. All this means that the percentage does not serve to the numerical comparison, for the reason that, for example, 3 goals in 10 matches does not guarantee 30 goals in 100 games, which corresponds precisely to 30%. We believe that the percentage serves to describe the proportion that each case occupies within the set, but it does not serve to compare each case among several sets. Later we will look for the solution of this problem of the numerical evaluation, typical of the Relative Frequency (FR) and the percentage.

But first, let'*s* look at the problem of another frequency also used in corpus linguistics in general. It is the Normalized Frequency (NF), which is calculated by the division of the Absolute Frequency (AF) by the Totality of Words (TW) counted in each section, multiplied by an appropriate Multiplier (M).

NF = AF / TW * M

For example, in the area of 1200 of the corpus there have been 7 736 words in total, which is the Totality of Words (TW). Then, the Normalized Frequency of <uoz> in 1200 is 3/7 736 * 100 000 = 38.8. We recommend using as Multiplier (M) the rounded number next to the Maximum of the base (TW): 96 059 (in the data set of 1400). We get to get the lower right table (NF):

AF	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
TW	7 736	36 052	40 957	64 999	96 059

NF.	1200	1250	1300	1350	1400
uoz	**38.8**	**22.2**	7.3	16.9	6.2
boz	0.0	8.3	19.5	27.7	36.4
voz	0.0	2.8	2.4	35.4	55.2

However, here in the Normalized Frequency (NF) there is also the same problem of lack of comparability characteristic of the data from different bases, and especially, of some rather small bases. We can not help but feel doubts about the NF figure of <uoz> in 1200, 3 among 7 736 whose NF is 38.8 in comparison with the NF in the same way <uoz> in 1250, 8 among 36 052 whose NF is 22.2. We wonder if 38.8 is really comparable with 22.2.

The essence of the problem is the same in both the Relative Frequency (FR) and the Normalized Frequency (NF) in the sense that the two calculate on different bases. Paradoxically, the two frequencies are used precisely when the bases are different, since if the bases are the same there is no need to resort to these frequencies and in the net Absolute Frequency we can make the numerical comparison without problem.

The problem of the lack of comparability discussed in this section is solved by the elimination of the set in question. In the example of the data of the three medieval forms, we would try to eliminate the set corresponding to 1200. It is the general practice in the statistical treatment. For example, in the sports world of baseball, players' scores are calculated with sufficient participation in the matches. Players who do not pass the established threshold are excluded from the assessment from the beginning. But we wonder what we do with the 1250 band, where frequencies are recorded within the base of almost a third of 1400 (37.5%).

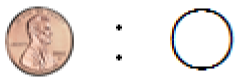
Our idea is to treat all data without distinction, but with common criteria of probability. Our method, which we will explain below, offers the evaluation of the data equally, with similar or distant bases, which shows the absolute robustness, in comparison with the traditional methods of the Relative Frequency (FR), including its variant Percentage, and the Normalized Frequency (NF), whose fragility we have seen in the cases of distant bases in this section.

To offer the solution to the problem of the Relative Frequency (FR) and the Normalized Frequency (NF), we present a new frequency formula. Our purpose is to look for a type of frequency, "Probabilistic Frequency" (PF), that represents the relative value of the Absolute Frequency (AF) within the set (base) with simple calculations of the probability (Ueda 2017). It is justified by the results of real experiments and by our intuition and thought that, for example, 30 out of 100 is more important (or significant) than 3 out of 10, even though both are equal of 0.3 (30%) in the probability of occurrence. To prove it, we resort to the binomial probabilistic formula. The way to get to know the Probabilistic Frequency (PF) goes through the following three steps: (1) Significance (S), (2) Expected Probability (EP) and, finally, (3) Multiplier (M).

Significance

We intuit and think that, for example, a player who has scored 28 goals in 100 games is more "important" and has contributed more to the team than the other who has scored 3 goals in 10 games, although the goal ratio of the first (28 / 100 = 28%) is less than that of the second (3/10 = 30%). To the degree of importance, we give the name of "Significance" (S). We start with some simple and special cases to reach the general case, applicable to frequencies in general.

To calculate Significance (S), we use binomial probability. To get to understand it, we start with a few examples as simple as a coin toss, where each face has the Expected Probability (EP) of 0.5 (50%).



The "face" of coin, which has the probability of 0.5 (50%) of appearing in a coin toss.

The following table shows the probabilities of two events (x = 0, 1) of a coin (trial number n = 1) of a coin: «face», with a value of 1, or «cross», with a value of 0. Expected Probability (EP) of each is 0.5, since there are two cases: face (1) or cross (0). In the table each event comes with its own Probability of Occurrence (PO), which we have just seen, the Cumulative Probability (CP), which accumulates with each corresponding Probability of Occurrence, and the Significance (S):

x	Caso	PO(x, 1, 0.5)	CP(x, 1, 0.5)	S(x, 1, 0.5)
cruz: x = 0	(0)	1/2 = 0.5	0.5	0
cara: x = 1	(1)	1/2 = 0.5	0.5 + 0.5 = 1.0	0.5

The Probability of Occurrence column PO (x) shows in the first row the probability of «cross» ($x = 0$), which is PO (0) = 1/2 = 0.5 and, in the second row, the probability of «face» ($x = 1$)) which is PO (1) = 1/2 = 0.5. The Cumulative Probability (CP) of $x = 0$, CP (0), is 0.5, which is equal to PO (0) and that of $x = 1$, CP (1), is 1.0, which is the sum of PO (0) = 0.5 and PO (1) = 0.5. The last Cumulative Probability (CP) is always 1.

Now, we define the "Significance" (S) of x, S(x, n, e), as corresponding to the Cumulative Probability of x - 1, CP (x-1, n, e):

S (x, n, e) = CP (x-$1, n, e$)

(x: Occurrence; n: Tests; e: Expected Probability)

The Significance (s) of $x = 0$, S (0), we define it as 0, because the Cumulative Probability of -1 does not exist:

S (0) = 0

The reason why we consider Significance as the Cumulative Probability of the Occurrence of the immediately preceding case is that the sum of the previous probabilities of the corresponding Occurrence is the probability of the significance of the numbers of previous occurrences. If we toss a coin, the Significance of the occurrence of 1 («face») is 0.5, which is complementary to the Risk (not «face», that is, «cross»), which is also 0.5. Therefore,

Significance + Risk = 1

which means that there is Significance of 0.5 (50%) of the appearance of the face and there is a Risk of 0.5 (50%). This means that if we bet on the appearance of the «face», there is a 50% risk, which we know and intuit without resorting to probability theory.

So far we have seen a very simple case in which we throw the coin only once. What happens if we throw the same coin twice? The following table shows the distribution of Occurrence Probabilities (OP) that they present in two tests of tossing a coin ($n = 2$). There are three possible cases: $x = 0, 1, 2$, that is, (0,0), (1,0) + (0,1) and (1,1):

x	Case	PO(x, 2, 0.5)	CP(x, 2, 0.5)	S(x, 2, 0.5)
$x = 0$	(0, 0)	1/4 = 0.25	0.25	0
$x = 1$	(0, 1); (1, 0)	2/4 = 0.50	0.25 + 0.50 = 0.75	0.25
$x = 2$	(1, 1)	1/4 = 0.25	0.75 + 0.25 = 1.00	0.75

This time the Expected Probability (EP) of «face» is equally 0.5. The Probability of Occurrence (PO) column shows that the PO of 0 face occurrences, PO (0), is 0.25 (cross, cross) = (0, 0), ie 1 of 4 cases. The total number of cases is 4, because 4 following cases are counted: (0, 0), (0, 1), (1, 0), (1, 1). The probability of 1 «face» occurrence, («face», «cross») + («cross», «face»); (1, 0) + (0, 1) is 0.5 (2 of 4 cases). And, finally, the probability of 2 «face» occurrences, («face», «face»); (1, 1), PO (2) is 0.25, which occurs 1 out of 4 cases. The Cumulative Probability (CP) column presents the summed probabilities from 0 to 2 in each occurrence: $x = 0, 1, 2$.

The last column of Significanc (S) corresponds to the previous case of Cumulative Probability (CP). The last Significance (S) of $n = 2$ is S (2) = 0.75, which represents a considerable increase with respect to the previous experiment in which the coin was thrown only once: 0.5 ($n = 1$), which means that 2 between 2 (S = 0.75) is much more "significant" (important) than 1 between 1 (S = 0.5), even though both are equal to 100% of Cumulative Probability (CP). However, Significance (S) still only reaches 0.75 (75%), which means that there is 0.25 (25%) Risk.

Now, we need the 3 parameters of the Significance (S): x: occurrences, n: total of the times of tests, and: Expected Probability (EP) in the form of the function S(x, n, e):

S(2, 2, 0.5) = CP (1, 2, 0.5) = 0.75

In the same way, the Significance (S) of $x = 1$ is:

S (1, 2, 0.5) = CP (0, 2, 0.5) = 0.25

Let'*s* see the experiment of 3 tests ($n = 3$):

x	Caso	PO(x, 3, 0.5)	CP(x, 3, 0.5)	S(x, 3, 0.5)
$x = 0$	(0,0,0)	1/8 = .125	.125	0
$x = 1$	(1,0,0), (0,1,0), (0, 0, 1)	3/8 = .375	.500	.125
$x = 2$	(1,1,0), (1,0,1), (0,1,1)	3/8 = .375	.875	.500
$x = 3$	(1,1,1)	1/8 = .125	1.000	.875

The Significance (S) of the last occurrence ($x = 3$) has increased in 0.875 and consequently now the Risk has decreased by 0.125 (1 - 0.875).

S(3, 3, 0.5) = CP(2, 3, 0.5) = 0.875 (87.5%)

If we bet that the «face» does not come out 3 times, there is a probability of 87.5% of winning the bet, which is the Significance (S); and the risk of losing the bet is 12.5%. We should increase the Significance (S) to at least 95% (0.95) and, if possible, up to 99%, with the Risks of 5% or 1%, respectively. In this way we lose the bet only 1 of 20 times, or 1 of 100 times.

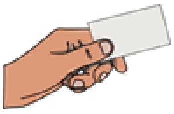
Let'*s* see the experiment of 10 trials (n = 10):

x	PO(x, 10, 0.5)	CP(x, 10, 0.5)	S(x, 10, 0.5)
$x = 0$	.001	.001	.000
$x = 1$	.010	.011	.001
$x = 2$	.044	.055	.011
$x = 3$	.117	.172	.055
$x = 4$	.205	.377	.172
$x = 5$	.246	.623	.377
$x = 6$	.205	.828	.623
$x = 7$	.117	.945	.828
$x = 8$	.044	.989	.945
$x = 9$	**.010**	**.999**	**.989**
$x = 10$	**.001**	**1.000**	**.999**

Finally when x = 9, we obtain the Significance S(9, 10, 0.5) = 0.989, higher than 95%, and the S(10) = 0.999, higher than 99%, which means that we can present the figure of 9 between 10 with Significance (S) greater than 95% and that of 10 between 10 with Significance (S) greater than 99%. Actually, when you toss the coin 10 times if the face of the coin comes out 9 times, the total probability of occurrences less than 9 [0, 1, 2, ..., 8] adds up to 98.9%, which is quite significant. That is to say, with the Significance of 98.9% we can affirm that 9 out of 10 is significant (important). It is significant or important in the sense that 9 or 10 out of 10 occur only with the probability of 0.010 + 0.001 = 0.011 (1.1%). In the same way, we can affirm that 10 out of 10 has the Significativity of 0.999 (99.9%). Compare with the cases of 1 between 1 (Significativity of 50%), 2 between 2 (75%) and 3 among 3 (87.5%)[10].

[10] Here it is not about the Probability of Occurrence (PO) but the Cumulative Probability (CP) of the cases until the immediately previous case. We observe the

So far we have seen the mathematical behavior of Significance (S), which depends on the three parameters: *x*: occurrences, *n*: total of the times of tests, and Expected Probability (EP). We have observed its movement according to *x* and *n*. Now let'*s* see what Significance (S) is presented according to the change in Expected Probability (*e*). The following table shows the Significance (*s*) of the occurrences (*x*) of events endowed with the Expected Probability (*e*) of 0.1, for example, the Expected Probability (*e*) of taking the "1" card out of the ten cards:


□□□□□□□□□□

x	PO(*x*, 10, 0.1)	CP(*x*, 10, 0.1)	S(*x*, 10, 0.1)
x = 0	.349	.349	.000
x = 1	.387	.736	.349
x = 2	.194	.930	.736
x = 3	.057	.987	.930
***x* = 4**	**.011**	**.998**	**.987**
***x* = 5**	**.001**	**1.000**	**.998**
x = 6	.000	1.000	1.000
x = 7	.000	1.000	1.000
x = 8	.000	1.000	1.000
x = 9	.000	1.000	1.000
x = 10	.000	1.000	1.000

For example, when *x* = 5, *n* = 10, *e* = 0.1, S(5, 10, 0.1) turns out to be 0.998, that is, the sum of the Occurrence Probabilities (PO) *x* = 0, 1, 2, 4 is 0.998. Therefore, when establishing the norm of Significance (S) in 0.99, 99% of the occurrences correspond to 0, 1, 2, 3, 4. It almost never appears 5 onwards (5, 6, 7, .. .) and there is a low probability of 0.01 (1%).

individual PO, for example, of a 5«faces» essay in 10 coins, it is only .246. When we add the odds of 0 to 4 faces, we get to the CP of .377. If we add the cases from 0 to 5 faces, we arrive at the CP of .623. Between .377 and .633 is the expected probability of .500. On the other hand, the reason why we add the cases of the beginning (0) to an immediately previous case (4) is that we use the complement of the CP (1 - CP) as an indicator of the degree of significance. For example, the 9-«faces» CP within 10 coins is .989 and its complement, .011 (1.1%). The probability of 1.1% is quite low so the null hypothesis that the currency is not biased is rejected.

We use the function of Excel BINOMDIST to obtain the Significance (S) of 0.349 in the cell of $x = 1$; $e = 0.1$:

$s = \mathrm{S}(x, n, e) = \mathrm{BINOMDIST}(x\text{-}1, n, e, 1)$
(s: Significance, x: Occurrence, n: Trials, e: Expected Probability)
$0.349 \leftarrow \mathrm{S}(1, 10, 0.1) = \mathrm{BINOMDIST}(0, 10, 0.1, 1)$

This means that when testing 10 times of the event with the Expected Probability (EP) of 0.1, the occurrence 1 ($x = 1$) corresponds to the Significance (S) of .349. The 2 occurrences ($x = 2$) of the same event correspond to 0.736:

Expected Probability

We have observed that the Significativity (s) is obtained by the function of S(x, n, e) or the function of Excel BINOMDIST:

$s = \mathrm{S}(x, n, e) = \mathrm{BINOMDIST}\ (x\text{-}1, n, e, 1)$
(x: Occurrences, n: Trials, e: Expected Probability)

By the function S(x, n, e) the Significance (s) is obtained by means of x: Occurrences, n: Trials, e: Expected Probability. However, in the practice of linguistic data analysis, unlike such coin toss or card draw experiments, the Expected Probability (EP) of the events from the beginning is generally not known. The parameters that are known are x: Occurrences, n: Trials (Sum) and the Significance (s) is established by the user. For this reason, we then elaborate the function E that returns the Expected Probability (EP) by means of x: Occurrences, n: Trials and s: Significance:

$e = \mathrm{E}(x, n, s) = \mathrm{E}(1, 10, 0.95)$

The function E (x, n, s) returns Expected Probability (EP) that is assumed from an event that occurs x times in n trials with Significance s. It is to presuppose the Expected Probability (e) of, for example, 5 occurrences ($x = 5$) in 10 trials ($n = 10$) with the Significance (s) of, for example, 0.99 ($s = 0.99$). With these three parameters, the Expected Probability (EP) is calculable.

Suppose we have had 2 successes ($x = 2$) in 10 experiments ($n = 10$). With these data, however, we can not expect 20 successes in 100 future experiments. Let'*s* see how the Expected Probabilities (e) are presented by increasing the number of experiments n = 10, 100, 1000, ...:

n	E(*n**0.2, *n*, 0.99)
n = 10	0.016
100	0.116
1,000	0.171
10,000	0.191
100,000	**0.197**
1,000,000	0.199
10,000,000	**0.200**
100,000,000	0.200
1,000,000,000	0.200

In the previous table with the condition that the Significance is 0.99 (99%), when obtaining 2 successes in 10 trials, its Expected Probability (*e*) is 0.016 (1.6%) and it is very far from the probability of success of 0.20 (twenty%). When *n* = 10,000 it reaches *e* = 0.191 (19.1%). From *n* = 10,000 onwards, the increase in Expected Probability (EP) is reduced. Finally we obtain *e* = 0.20 (20%) upon arriving at *n* = 10 000 000. This characteristic of the Expected Probability (EP) is important, since by it we can know what theoretical probability exists in each case of 2 among 10, 20 among 100 , 200 among 1000, so on. We draw attention especially the first cases where the magnitude of the base 10, 100, 100 is reduced, which causes the little expected probability (0.016, 0.114, 0.171).

We are going to carry out the same experiment changing the expected probability (*e*) in 0.100, 0.200, ..., 1.000.

BinE	*n* = 10	100	1 000	10 000	100 000	1 000 000
g = 0.10	0.001	0.042	0.079	0.093	0.098	0.099
0.20	0.016	0.116	0.171	0.191	0.197	0.199
0.30	0.048	0.198	0.267	0.289	0.297	0.299
0.40	0.093	0.287	0.364	0.389	0.396	0.399
0.50	0.150	0.381	0.463	0.488	0.496	0.499
0.60	0.218	0.479	0.563	0.589	0.596	0.599
0.70	0.297	0.582	0.665	0.689	0.697	0.699
0.80	0.388	0.691	0.769	0.791	0.797	0.799
0.90	0.496	0.809	0.876	0.893	0.898	0.899
1.00	0.631	0.955	0.995	1.000	1.000	1.000

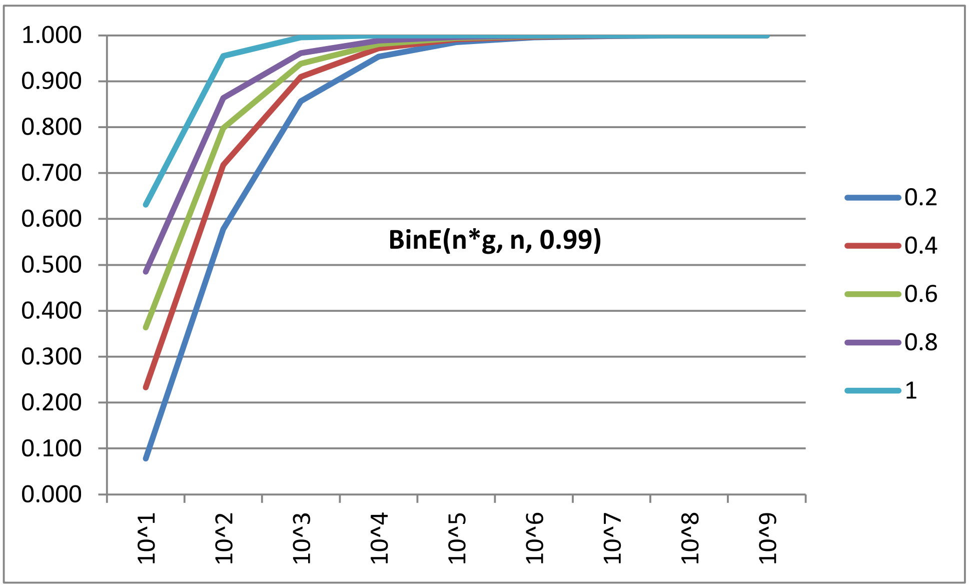
Expected probability *e* = BinE(*n*g, n,* 0.99)

In the previous table, we observe that the number of trials (*n*) is reduced to reach the goal of 20%, 40%, ..., 100%, according to the increase in the success ratio (*g*). Therefore, we look for the target compliance ratio by dividing the expected probability (*e*) by the success rate (*g*):

BinE	*n* = 10	100	1 000	10 000	100 000	1 000 000
g = 0.10	1.0%	42.4%	79.1%	93.1%	**97.8%**	**99.3%**
0.20	7.8%	57.8%	85.7%	**95.4%**	**98.5%**	**99.5%**
0.30	15.8%	66.1%	88.9%	**96.5%**	**98.9%**	**99.6%**
0.40	23.3%	71.8%	91.0%	**97.1%**	**99.1%**	**99.7%**
0.50	30.1%	76.1%	92.6%	**97.7%**	**99.3%**	**99.8%**
0.60	36.4%	79.8%	93.9%	**98.1%**	**99.4%**	**99.8%**
0.70	42.4%	83.1%	**95.0%**	**98.5%**	**99.5%**	**99.8%**
0.80	48.5%	86.3%	**96.1%**	**98.8%**	**99.6%**	**99.9%**
0.90	55.1%	89.9%	**97.3%**	**99.2%**	**99.8%**	**99.9%**
1.00	63.1%	**95.5%**	**99.5%**	**100.0%**	**100.0%**	**100.0%**

Target compliance ratio = BinE(*n***g*, *n*, 0.99) / *g*

For example, the target compliance ratio of the case of [*n* = 10: *g* = 0.200] is 0.016 / 0.200 = 0.078 (7.8%). This means that two successes within 10 trials, if 99% of significance is desired, is not assured 20%, but 1.6% (0.016), which reaches only 7.8% of the goal of 0.2000. To achieve 95% (0.950) of the target compliance ratio, approximately 10,000 trials are needed. The objective compliance ratio (BinE / *g*) varies according to *g* and the more the success ratio (*g*), in the fewer tests the compliance ratio is reached, for example, 95%. For example, when *g* = 0.8, in 1,000 trials the compliance ratio reaches 96.1% (0.961).



In the previous table and the graph, we take into account that when $g = 1$ (100%), we reach 100% of the compliance rate in 1000 trials, while when $g = 0.2$ (20%), we need 1000 000 000 tests. The compliance ratio of 95% is achieved, instead of 100%, when $g = 0.2$, 0.4, 0.6, and $n = 10{,}000$. In any case, if g is reduced, a large amount of tests (n) is needed to give sufficient security (99%).

Therefore, when dealing with data of less than 1000 (n), especially with the reduced success rate (g), we must be careful in the handling of the relative and normalized frequency. The linguistic data is usually of low probability (g), for example the frequency of words or morphemes in less than 1% (0.01) of the whole. In this case we recommend using the expected probability, the basis of the probabilistic frequency, which is usually smaller than the success rate, but always offers the guaranteed significance of, for example, 99%.

Probabilistic Frequency

We try to calculate the "Probabilistic Frequency" (PF) in the form of Expected Probability (e) * Multiplier (m):

PF = $e * m$ (e: Expected Probability, m: Multiplier)

The Probabilistic Frequency (PF) is obtained by the function of the Expected Probability E (x, n, s) in combination with the Multiplier (m).

PF = $e * m$ = E (x, n, s) * m

21.5 = E (3, 3, 0.99) * 100

It is convenient that the amount of the Multiplier (m) be of the similar magnitude of the Maximum of Sum or Total of Words in rounded form.

The lower left table shows the Absolute Frequency (AF) and Vertical Sum and the right table is the Probabilistic Frequency (PF) with the multiplier (m) = 100. Now the Probabilistic Frequency (PF) of [uoz: 1200] reaches the figure of 21.5, unlike the Relative Frequency (AF) of 3 between 3: 3/3 * 100 = 100.0, which is incomparable with, for example [uoz: 1250], AF = 8/12 * 100 = 66.7. It turns out that the Relative Frequency (FR) of 3 among 3 (1,000) is higher than that of 8 among 12 (.667), while the Probabilistic Frequency (PF) of 3 among 3 is 21.5 and that of 8 among 12 is 30.2, which shows the greater importance of 8 among 12.

AF	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Suma	3	12	12	52	94

PF:100	1200	1250	1300	1350	1400
uoz	21.5	30.2	3.9	9.7	1.9
boz	.0	3.9	30.2	20.0	25.9
voz	.0	.1	.1	28.2	43.9

We have seen that the Relative Frequency (FR) is not appropriate to perform the comparative evaluation of the figures. Instead, we have introduced the Probabilistic Frequency (PF) based on the Sum of the compared forms. Now it is the Probabilistic Frequency (PF) calculated with the multiplier 100, with which the Significativity of .99 (99%) has been used, enough to be quite reliable. However, we are surprised by the magnitude of 21.5 in the case of 3 among 3 and 30.2 in the case of 8 among 12. They are correct within the Significance of .99 (99%), which is conditioned by the amount of the Multiplier (100). We think that this is due to the adoption of Sum of the frequencies of the forms in question as a basis for comparison. Let'*s* see the possibility of the other base, also very usual in corpus linguistic studies, the total number of words or letters.

The lower left table is the Absolute Frequency (AF) and the total number of words (TW):

AF	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
TW	7 736	36 052	40 957	64 999	96 059

PF:10^5	1200	1250	1300	1350	1400
uoz	5.6	8.1	1.1	7.3	1.9
boz	.0	1.2	7.1	14.8	23.7
voz	.0	.0	.0	20.5	39.1

The Normalized Frequency (NF), which in corpus linguistic studies is usually used with the total number of words (TW) in the form of, for example, 3 / 7736 * 100 000 = 38.8 in [uoz: 1200] presupposes that 3 among 7736 corresponds to 38.8 among 100 000, by the ratio formula:

3/7 736 = 38.8 / 100 000

However, from a probabilistic point of view this presupposition is not reliable, as well as the assumption that 3 successes in 10 trials correspond to 30 successes in 100 trials, which is the basis of the percentage, as we have seen previously. In the practice of comparison of figures, the Probability Frequency PF is more reliable, with which a gradual displacement of <uoz> (1200-1250) is seen by <boz> (1300) to <voice> (1350-1400). The same observation is possible with the Probabilistic Frequency (PF) of Sum with the multiplier 100. However, the table of the Probabilistic Frequency (PF) with the total number of words (TW) gives the most realistic figures.

Programs

Program-1: BinS (Excel VBA)

```
Function BinS(x, n, e)
'Significance s (x: occurrence, n: trials, e: expectation probability)
  If x = 0 Then BinS = 0: Exit Function
  BinS = Application.BinomDist(x - 1, n, e, 1)
End Function
```

Program-2: BinE (Excel VBA)

```
Function BinE(x, n, s)
'Expectative probability e (x: occurrence, n: trials, s: significance)
  Dim i, k, r, mn, mx, sc: If x = 0 Then BinE = 0: Exit Function
  r = 10 ^ 6: mx = r 'precision: maximum search
  Do While k < 1000
      i = (mx + mn) / 2 'midpoint between maximum and minimum
      BinE = i / r 'candidate of expectative probability
      sc = BinS(x, n, BinE) 'the candidate's own significance
      If sc < s - 1 / r Then 'If sc does not reach s-1 / r ...
          mx = i 'lower the search maximum to the midpoint
      ElseIf sc > s + 1 / r Then 'Si sc sobrepasa a s-1/r...
          mn = i 'raise the minimum to the midpoint
```

```
            Else 'If sc falls within the scope of s ± 1 / r ...
                  Exit Do 'leave the loop
            End If
            k = k + 1
      Loop
End Function
```