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 observed in space and time

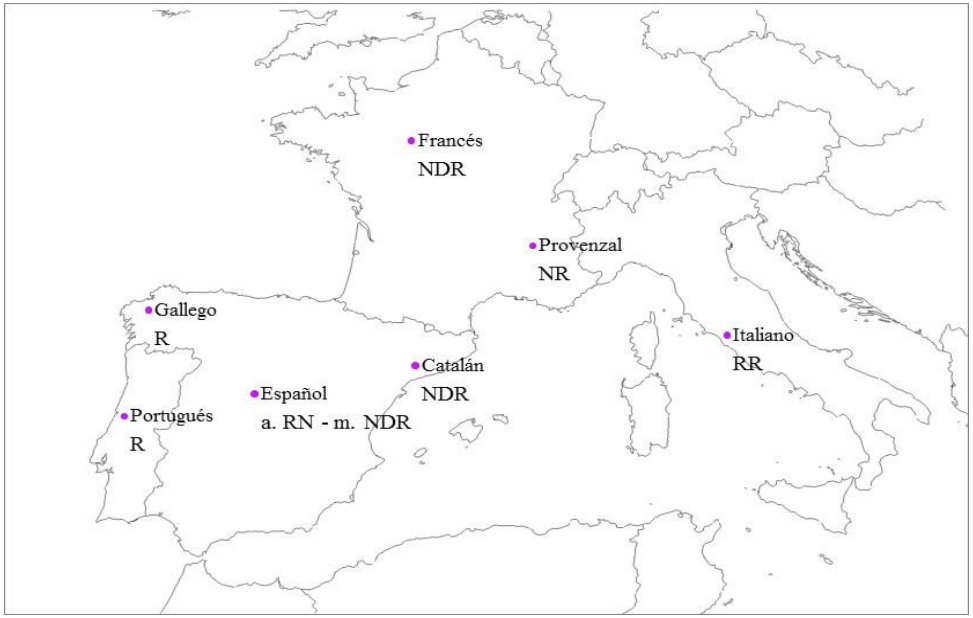
Hiroto Ueda, University of Tokyo

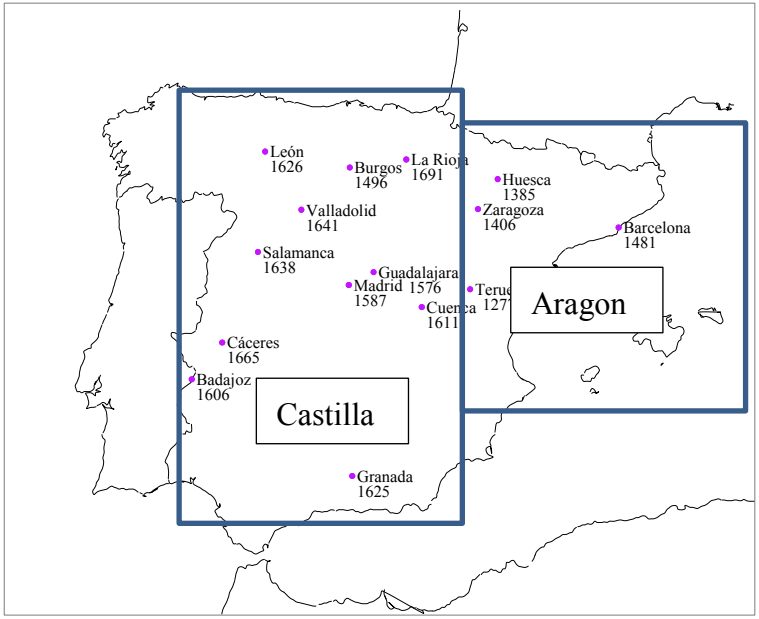
1. Introduction

Frequency and dispersion of linguistic forms in one text

Spanish texts of the Middle and Modern Ages

Castile and Aragon with extension to Navarra

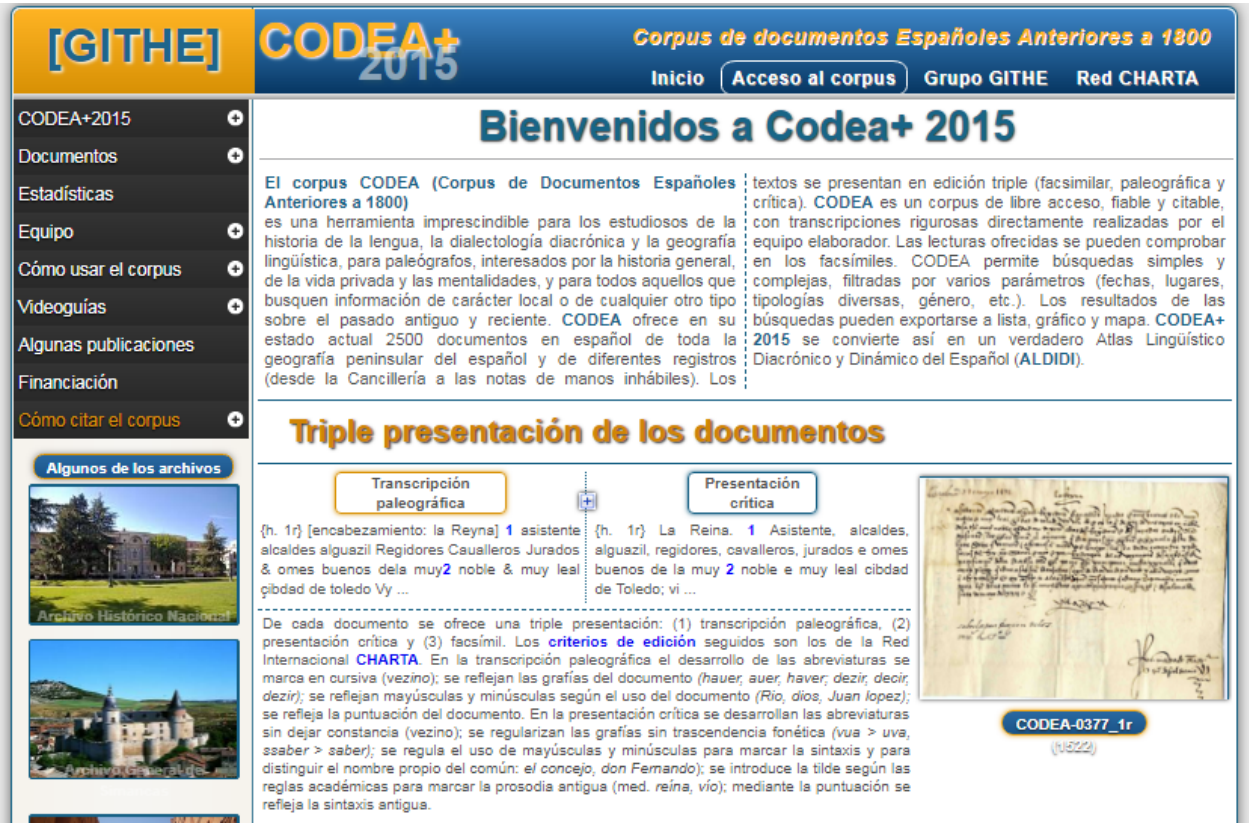




Irregular forms of future of verbs: *poner* 'to put', *tener* 'to have' and *venir* 'to come'

Corpus: CODEA «*Corpus de documentos Españoles Anteriores a 1800*»

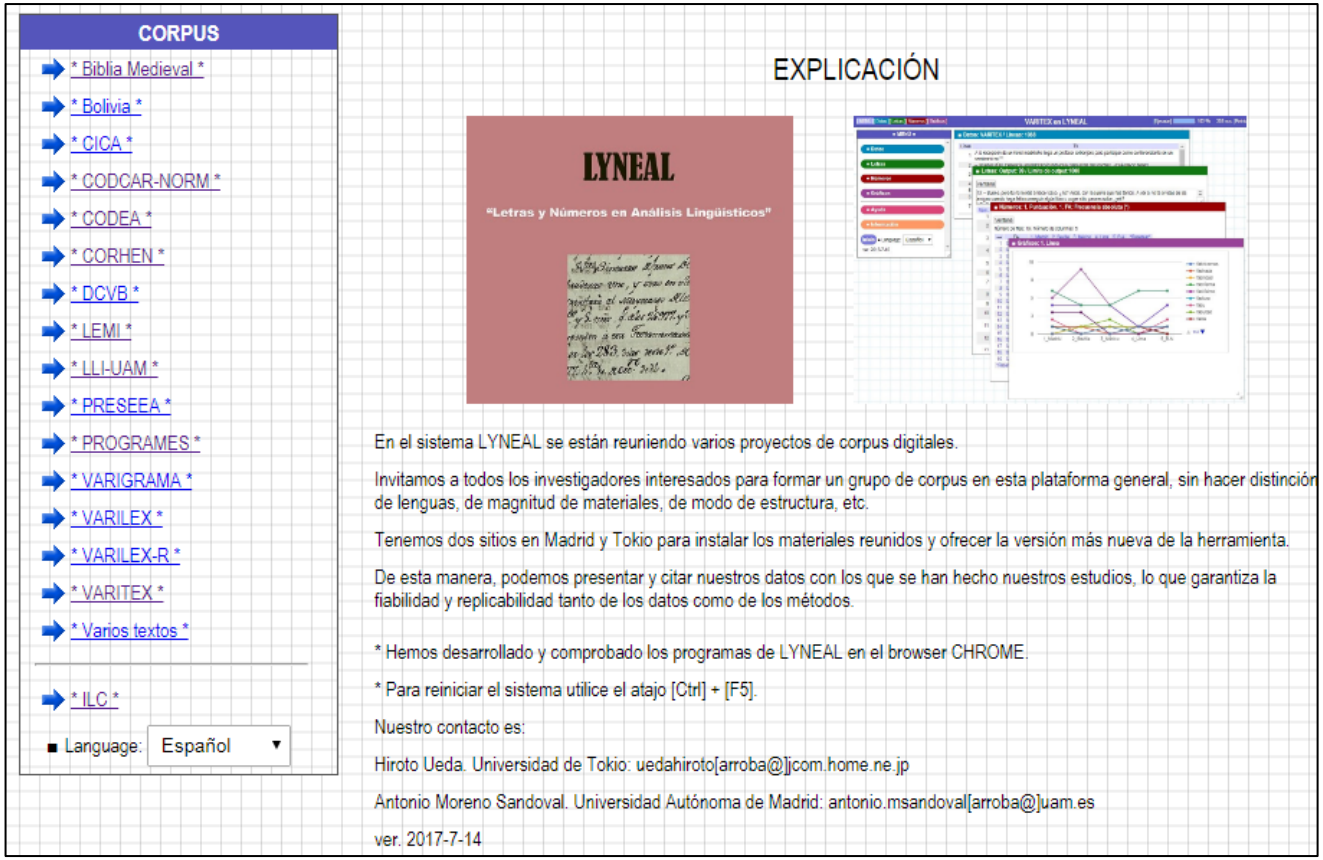
('Corpus of Spanish documents before 1800')



http://corpuscodea.es/

Analizer: LYNEAL («*Letras y Números en Análisis Lingüísticos*»

('Letters and Numbers in Linguistic Analysis').



http://shimoda.lllf.uam.es/ueda/lyneal/

2. Frequency and dispersion

2. 1. Probabilistic frequency

Absolute Frequency (AF)
Relative Frequency (RF)
Normalized Frequency (NF)
Probabilistic Frequency (PF)

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

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

3 / 12 = 25 / 100 = 25 % (?)

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

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

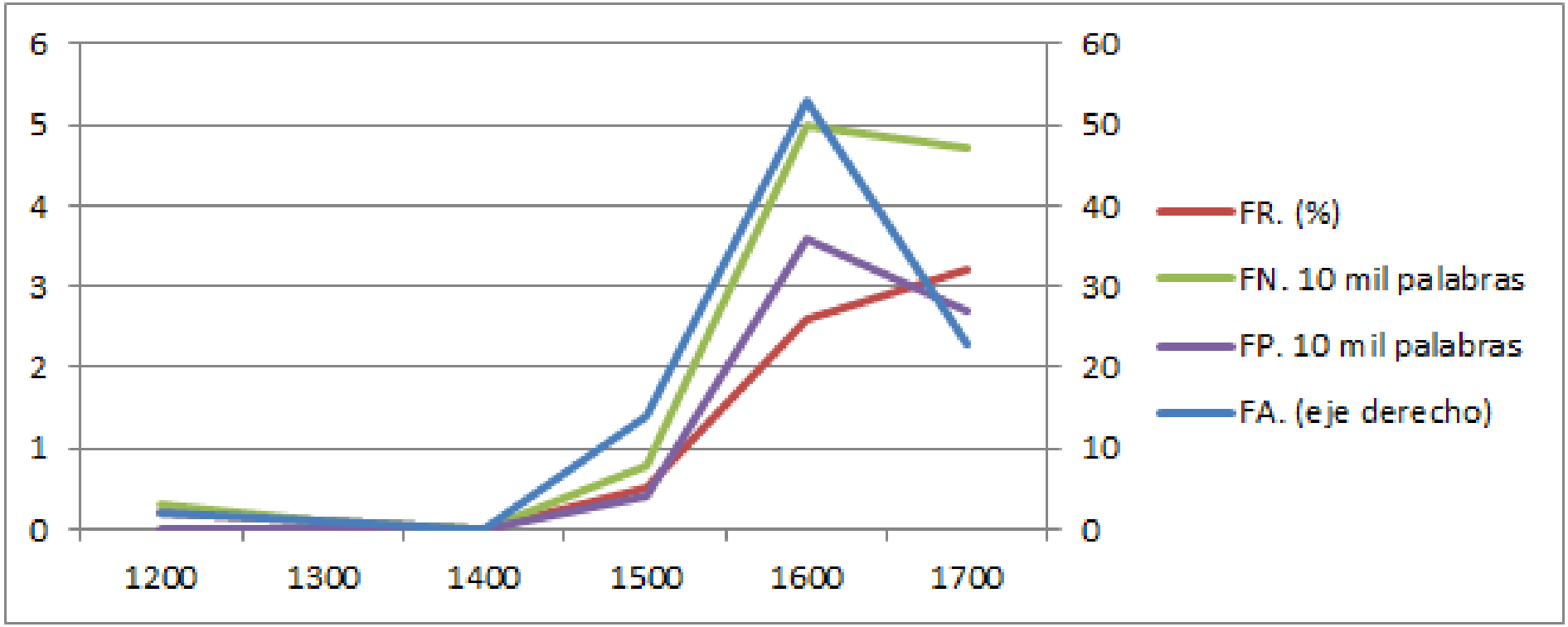
Probabilistic Frequency (PF):

Significance (*s*) of the *x* successes in *n* trials of an event endowed with the expectative binomial probability (*e*)

s = BinS(x, n, e) = Excel.BINOMDIST(x-1, n, e, 1)

e = BinE(x, n, s)

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



PF (Probabilistic Frequency): Significance of 99%

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.

2. 2. Block dispersion

Standard Deviation (SD):

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

where M is mean, and N is the total number of data.

Coefficient of Variation (CV):

$$CV = SD / M$$

Normalized Standard Deviation (NSD):

$$NSD = SD / SD.max$$

How to find the SD.max:

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

Eg. (10, 0, 0, 0, 0) → Maximum value of Standard Deviation (SD.max)

(K, 0, 0, ..., 0)

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

$$K = Sum = N\, M$$

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

$$= \{[(M\, (N - 1))^2 + M^2\, (N - 1)] / N\}^{1/2} \quad \leftarrow \quad M \text{ al exterior}$$

$$= \{[\underline{M^2\, (N - 1)^2} + \underline{M^2\, (N - 1)}] / N\}^{1/2} \quad \leftarrow \quad M^2 \text{ es común}$$

$$= \{M^2\, (N - 1)\, [\underline{(N - 1) + 1}] / N\}^{1/2} \quad \leftarrow \quad M^2\, (N - 1) \text{ es común}$$

$$= \{(M^2 (N - 1) \underline{N / N}\}^{1/2} \quad \leftarrow \quad (N - 1) + 1 = N \quad \leftarrow (N-1) + 1 = N$$

$$= [(M^2 (N - 1)]^{1/2} \quad \leftarrow \quad N / N = 1$$

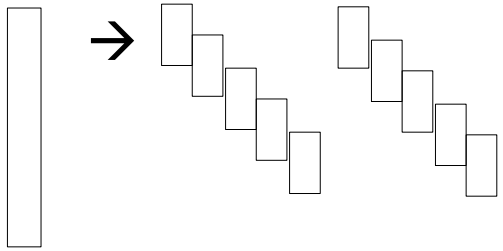
$$= M (N - 1)^{1/2} \quad \leftarrow \quad (M^2)^{1/2} = M$$

Normalized Standard Deviation (NSD):

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

Block Dispersion:

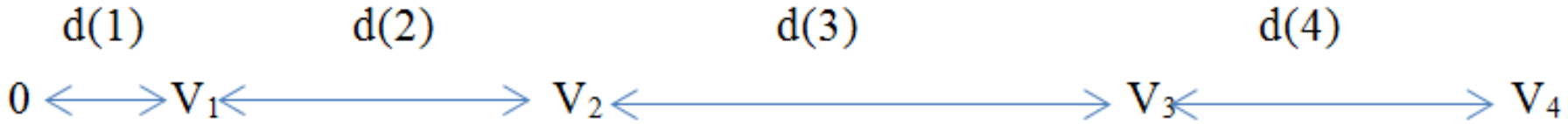
A: □ □ □ □ □ □ □ □ □ □

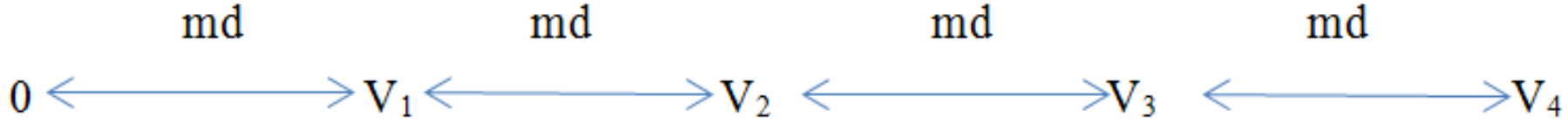
B: 

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

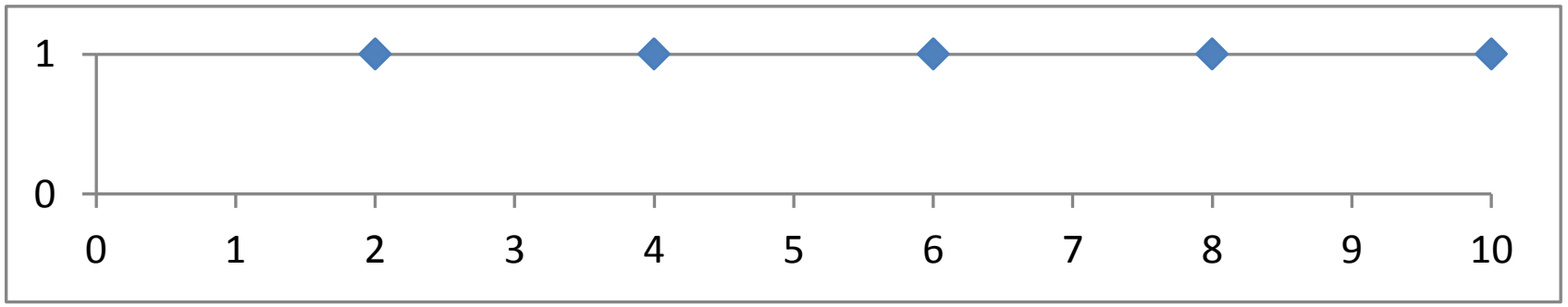
Disp.b. = 1 - NSD.b

2. 3. Uniformity



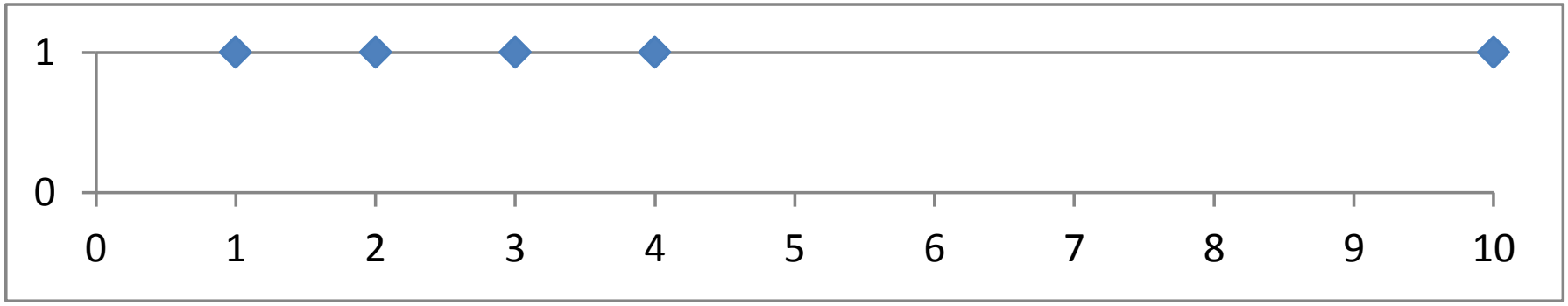


Element	Position	Distance	Mean distance	|Dis. -M. Dis.|
x1	2	2	2	0
x2	4	2	2	0
x3	6	2	2	0
x4	8	2	2	0
x5: M	10	2	2	0

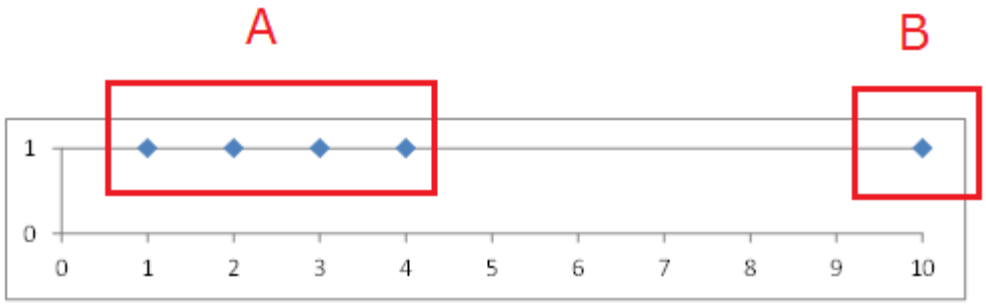


Position (2, 4, 6, 8, 10), Coagulation: 0.000, Uniformity: 1,000

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



Maximum of the sum of the distances (S.Dist.max) in A :

$$S.Dist.max\ (A) = (5 - 1) * | 1 - 10 / 5 | = 4 * | 1 - 2 | = 4$$

$$S.Dist.max(A) = (N - 1) * |1 - M / N|$$
$$= (N - 1) * |N - M| / N$$
$$= (N - 1) * (M - N) / N \quad \leftarrow M \geq N$$

Maximum of the sum of the distances (S.Dist.max) in B :

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

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

Maximum of the total sum of the distances (S.Dist.max) is:

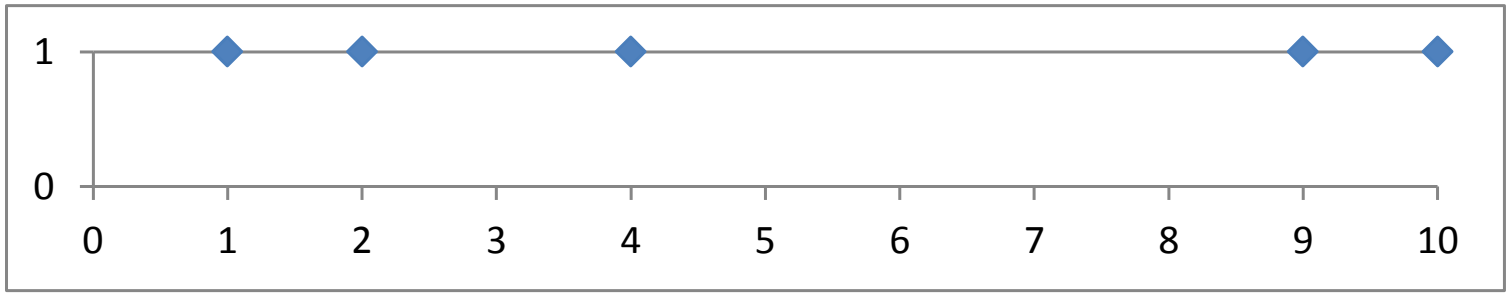
$$S.Dist.max = S.Dist.max\ (A) + S.Dist.max\ (B)$$
$$= 2 * (N - 1) * (M - N) / N = 4 * 2 = 8$$

$$Coagulation = S.Dist / S.Dist.max$$

$$Uniformity = 1 - Coagulation = 1 - S.Dist / S.Dist.max$$

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

Coagulation = 6 / 8 = .750; Uniformity = 1 - .750 = .250:



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

2. 4. Stable use

Juilland and Chang Rodríguez (1964), Usage (U):

$U = F * Disp$, F: Frequency (F), Disp: Dispersion

$Disp = 1 - SD / (2 * Mean)$

$Disp = 1 - NSD$

Stable Usage (SU):

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

3. Application

3. 1. Epenthetic forms

tener 'to have', *poner* 'to put', *venir* 'to come': *tenrá, terné, tendré, tendría, venr(r)án, vernié* etc.

Lloyd (1987: 496-7) / Penny (2006: 242):

"reinforcement" (*venr(r)án*) → NR

"metathesis" (*terné, vernie*) → RN

"epenthesis" (*tendrá*) → NDR

"assimilation" (*porrá, verrán*) → RR

Moreno Bernal (2004: 155) "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).

Marina Serrano (2018), CODEA

Aragon, Teruel (1277), Huesca (1285) and Zaragoza (1406). Barcelona (1481)

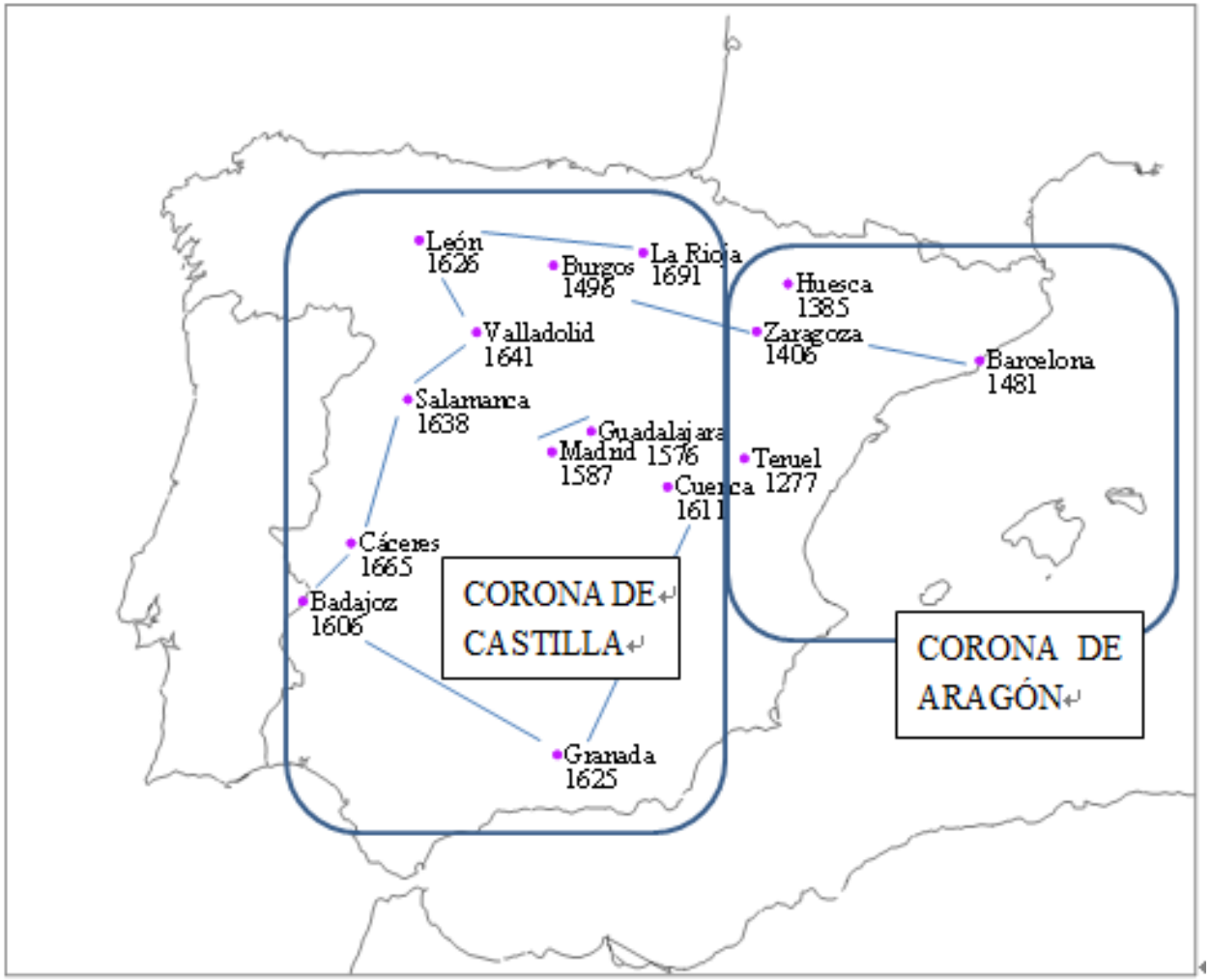


Fig. 3.1. First appearance of the NDR form

3. 2. Aragon and Navarra

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

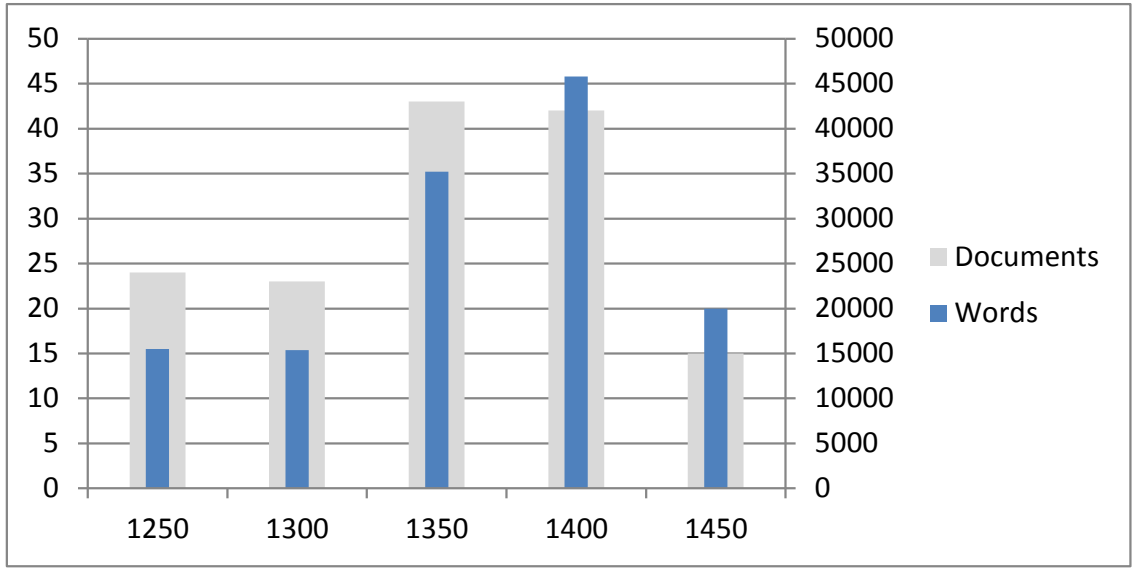


Fig. 4.2.a. Number of documents and words in Aragón and Navarra

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

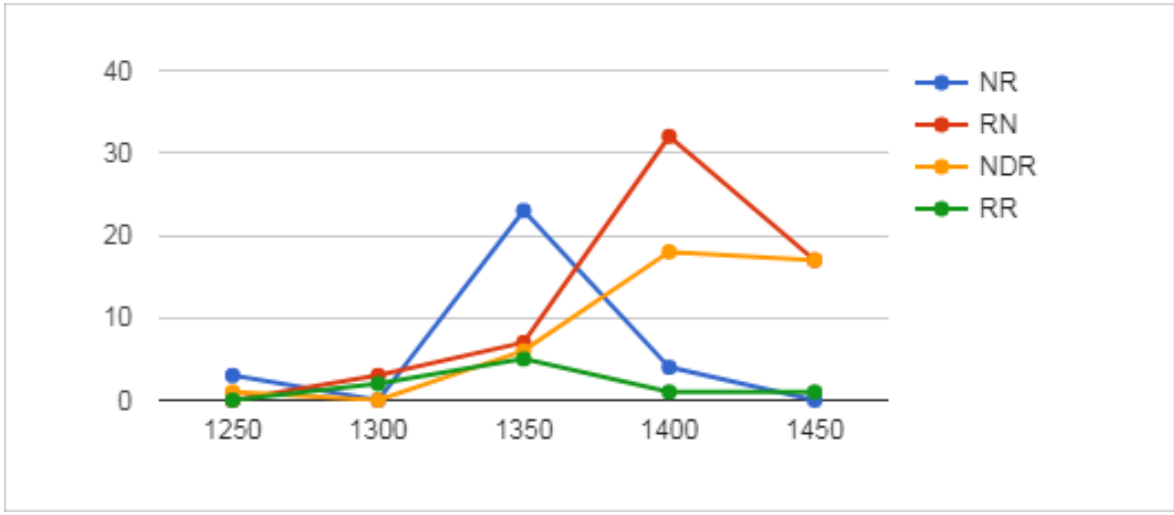


Fig. 4.2.c. Future forms in NR, RN, NDR, RR in Aragón. Absolute frecuency

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

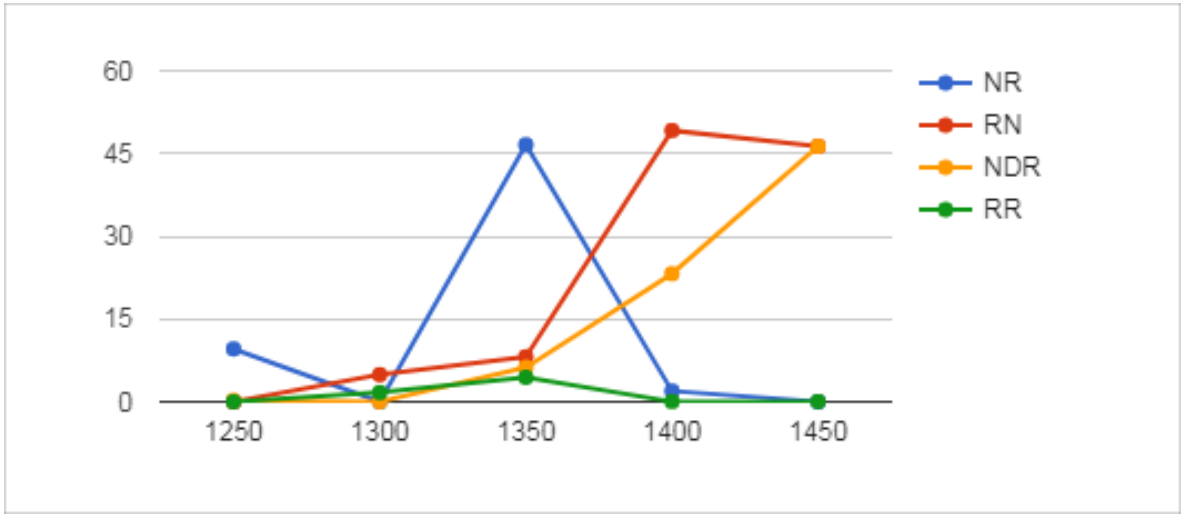
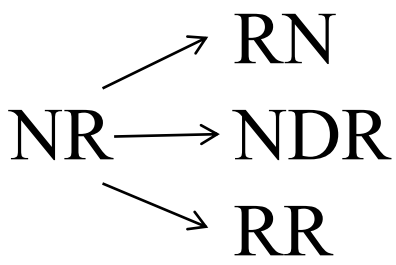


Fig. 4.2.d. Future forms in NR, RN, NDR, RR in Aragón

Probabilistic frequency for 100 thousand words



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 Castilla 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

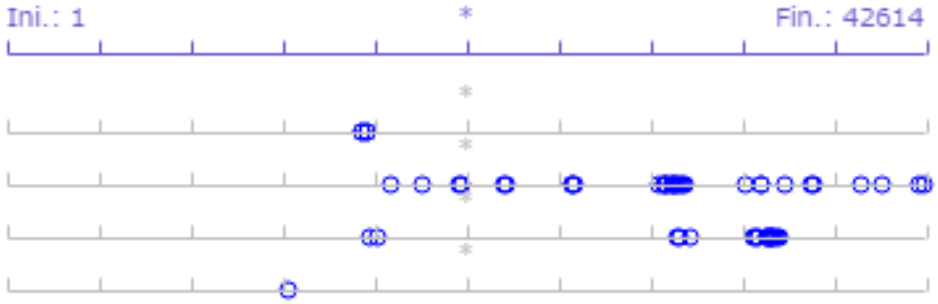


Fig. 4.2.f. Frequency and uniformity of RN, NDR, RR in Castilla of 1400

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

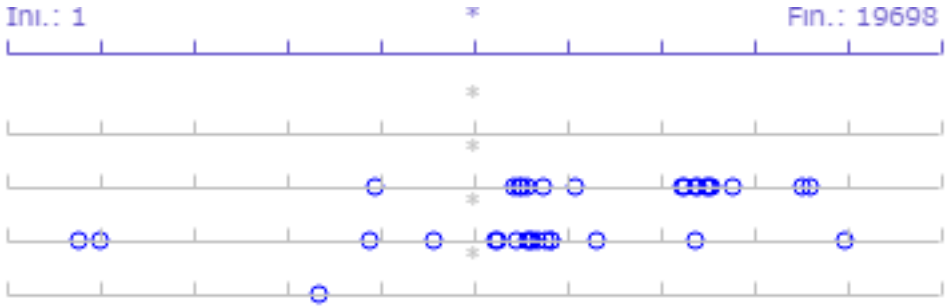


Fig. 4.2.g. Frequency and uniformity of RN, NDR, RR in Castilla of 1400

3. 3. Castile

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

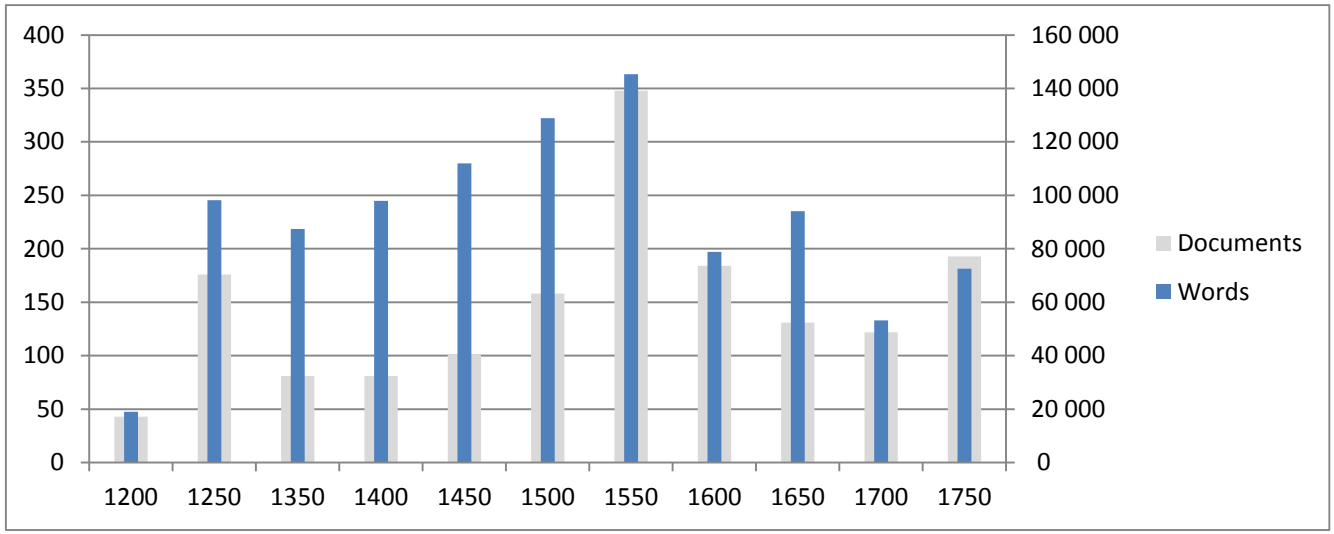


Fig. 4.3.a. Number of documents and words in Castile

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

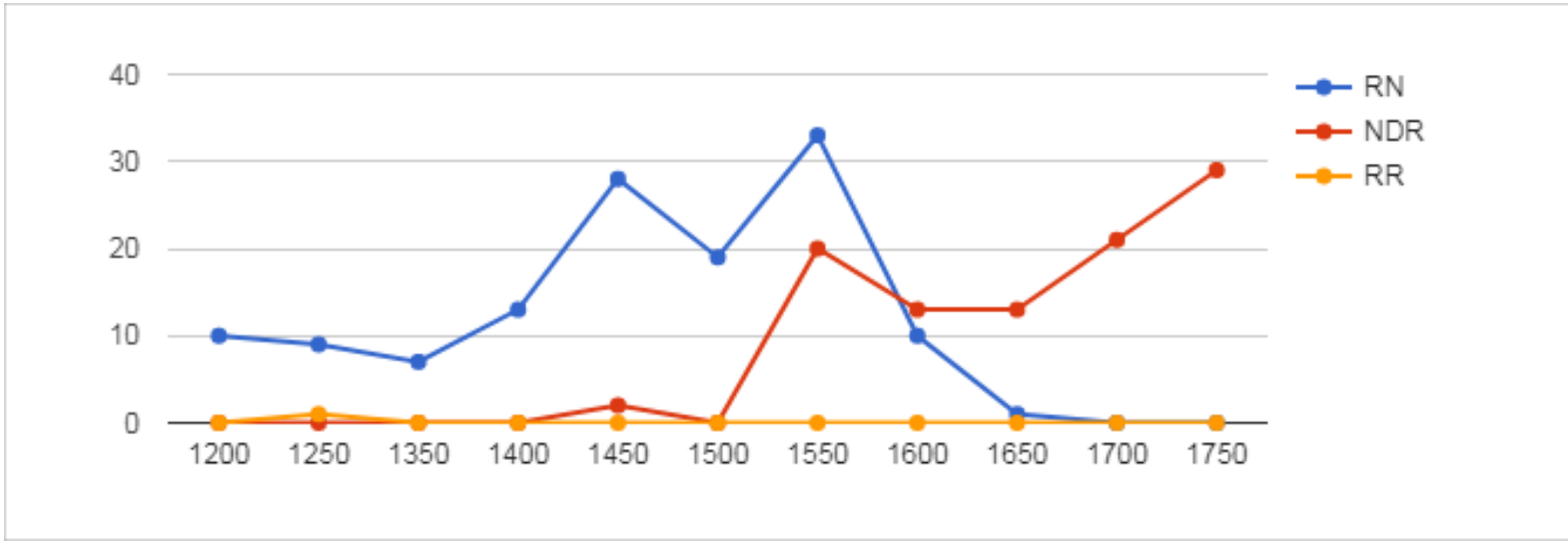


Fig. 4.3.b. Future forms in RN, NDR, RR in Castile. Absolute frecuency

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

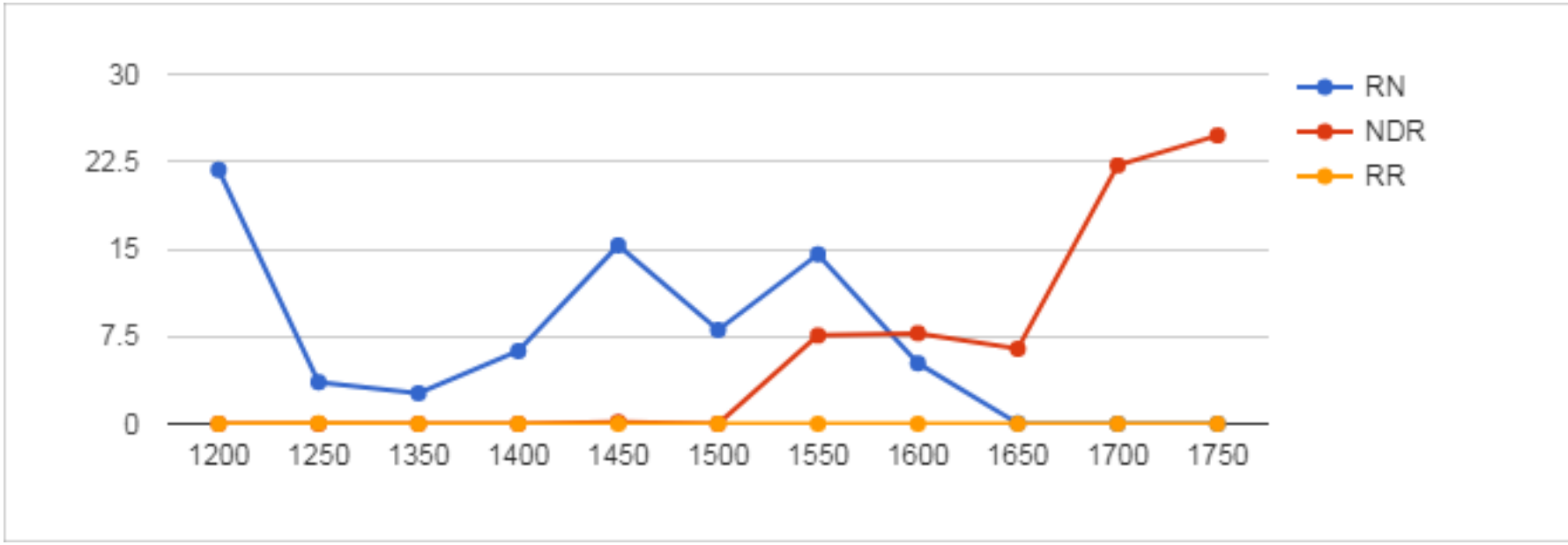


Fig. 4.3.c. Future forms of conjugation in RN, NDR, RR in Castilla

Probabilistic frequency per 100,000 words

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

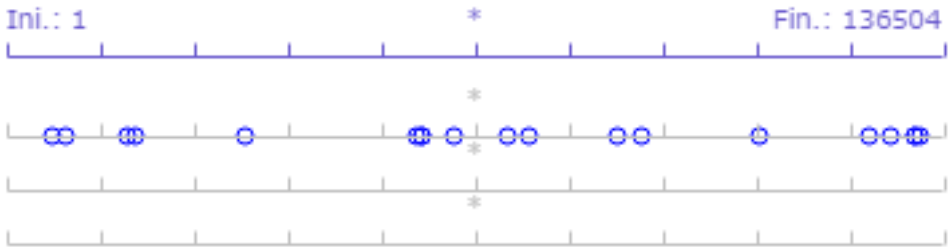


Fig. 4.3.d. Frequency and dispersion of RN, NDR, RR in Castilla de 1500

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 Castilla of 1550

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

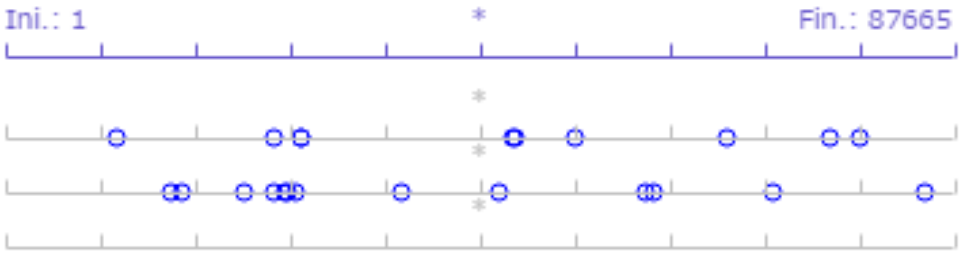


Fig. 4.3.f. Frequency and dispersion of RN, NDR, RR in Castilla of 1600

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

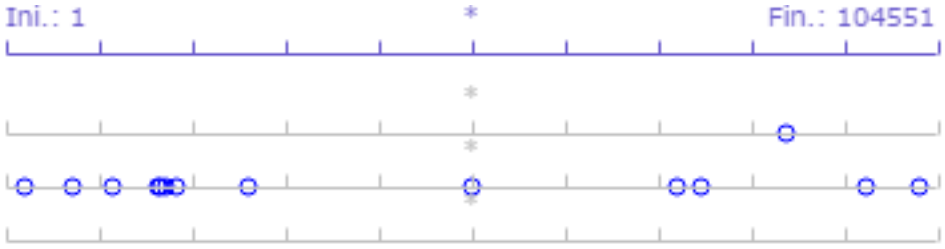


Fig. 4.3.g. Frequency and dispersion of RN, NDR, RR in Castilla of 1650

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

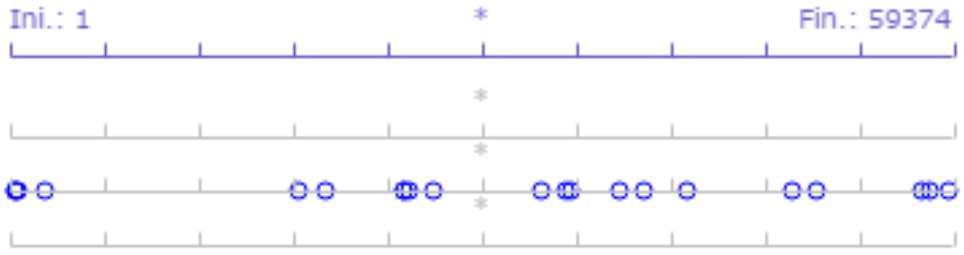


Fig. 4.3.h. Frequency and dispersion of RN, NDR, RR in Castilla of 1650

4. Discussion

Aragon in 1400: NR → NDR, for example, *tenrá, tenrán, venrán*, ('he will have', 'they will have', 'they will come')

Castile in 1600: RN → NDR, for example, *terné* en *tendré*

[n] [d] [r]

Lausberg (1973: 379):

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.*

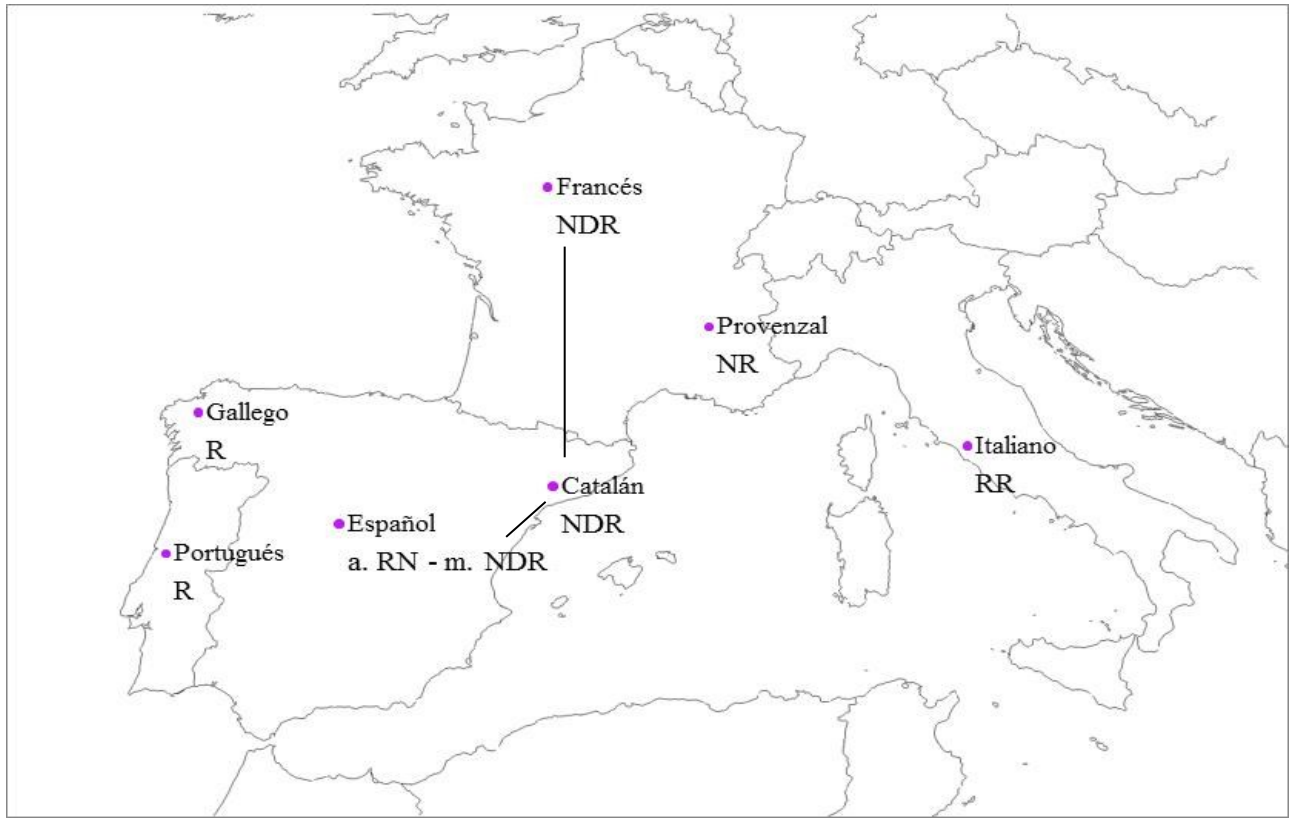
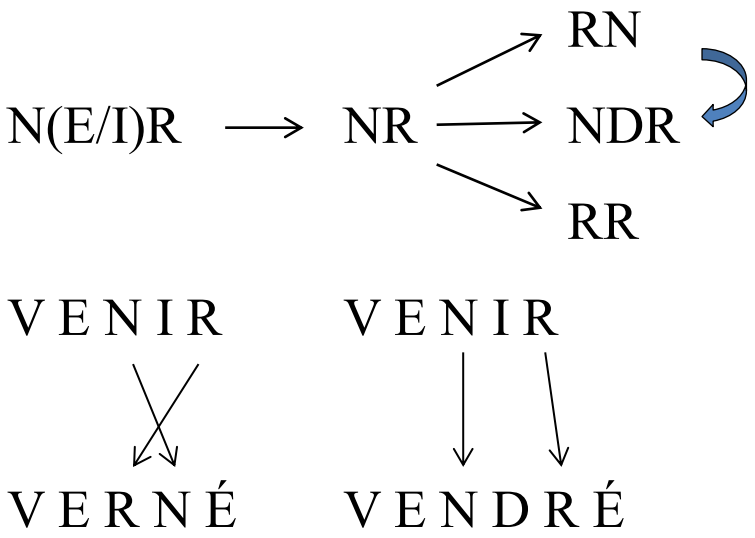


Fig. 5. Future forms with R, RN, NDR, NR, RR in Western Romania



Alvar and Pottier (1983: 251):

"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.)".

→ *vend-ré* with epenthesis is "more uniform" with the lexeme *venir*, than *verné* with metathesis.

Peny (2006: 111):

Epenthesis: 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')

Metathesis: N'R > RN (GEN(E)RU > *yerno* 'son-in-law', VEN(E)RIS > *viernes* 'Friday')

Menéndez Pidal (1972: 160-161):

Epenthesis: "Grupos interiores romances" (Romance interior groups)

Metathesis: "Cambios fonéticos esporádicos" (Sporadic phonetic changes)

The epenthesis is general, while the metathesis is special.

5. Conclusion

Hanssen (1913: 119):

"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 Aragón and the lexical transfer from RN to NDR in Castile.

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.

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 bunpo. (Gramática histórica de las lenguas románica.)* Tokio. Asahisyuppansha.

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 doctora 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 bunpoo handobukku. (Manual de la grama/tica española para estudiantes japoneses).* Tokio. Kenkyusha.

_____. 2017. *Análisis de datos cuantitativos para estudios lingüísticos.* https://lecture.ecc.u-tokyo.ac.jp/~cueda/gengó4-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.

Appendix

Programa-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
```

Programa-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
```