En comptes de contar contes, comptem les llengües dels contes.

Institut Pau Vila

Planter de sondeigs i experiments 2024

3r-4t d'ESO planter_84

Alumnes: Roger Forrellat, David Sánchez, Martina Porcar, Pol Muñoz i Pedro Couto

Professora: Lucia Bayo Delgado

ÍNDEX

1. ABSTRACT

The target of this project is to detect the language of a text through a spreadsheet, using the parameters we have found with our research: counting words and characters, making percentages of vowels, consonants and word length, and also making means of words and characters lengths. With these parameters, we have created a detector of languages, which gives a correct result or sometimes does not, depending on the kind of text and its length.

2. INTRODUCCIÓ

Sempre hem pogut saber quin tipus d'idioma té un text o un altre mitjançant les grafies o altres indicacions que reconeixem. Llavors, ens hem encuriosit sobre com podem nosaltres detectar una llengua amb altres factors i paràmetres que trobem mitjançant l'estudi estadístic de diferents textos. Per enfocar la nostra idea, ens hem centrat en les següents cinc llengües: català, castellà, francès, alemany i anglès, perquè són les que més ens envolten i hi tenim més interès.

3. OBJECTIUS

El nostre objectiu és, a partir de textos i per a cadascuna de les cinc llengües, crear amb un full de càlcul un perfil estadístic que pugui caracteritzar cada llengua, de manera que davant d'un text nou el full de càlcul pugui detectar l'idioma en què està escrit.

Les característiques que mirem són: llargada de paraules, quantitat de caràcters i paraules, quantitat de consonants i vocals i mitjana de caràcters per paraula. Introduïm diferents textos amb les cinc llengües disponibles i recollim totes les dades de tots els textos.

4. HIPÒTESIS

Els indicadors que pensem a priori ens podran donar la diferenciació de llengües són:

- L'anglès és l'idioma amb menys caràcters per paraula o amb paraules més curtes.
- L'alemany és l'idioma amb paraules més llargues.
- L'espanyol és l'idioma que més paraules necessita per expressar una idea.
- El català és l'idioma que més vocals té.
- El francès té gran presència de consonants com la s i la x.

En aquest treball és més important l'objectiu (detectar l'idioma d'un text) que contrastar les hipòtesis inicials. Les que hem formulat han estat un punt de partida per començar el treball, però quan hem arribat a l'anàlisi ens hem adonat que el que importava era fixar noves hipòtesis sobre criteris útils a partir dels resultats que anàvem obtenint.

➢ Respecte a la detecció:

Pensem que no detectarà amb precisió textos com la Declaració Universal dels Drets Humans, perquè és un text expositiu que té un lèxic prou similar en la majoria dels idiomes, ja que fa servir paraules que venen del llatí i, per tant, les freqüències de les lletres seran prou semblants.

5. RECOLLIDA DE DADES

POBLACIONS I MOSTRES

- Poblacions: caràcters, paraules i frases de tots els textos escrits en les cinc llengües.
- Mostres per recopilar informació: Totes les tipologies són narratives, des d'infantils fins a adults. Hem agafat les paraules i caràcters de quatre llibres: Harry Potter i la pedra filosofal, el Quixot, la Revolta dels animals i El petit príncep.
    - Hem descartat altres mostres narratives, perquè les traduccions pertanyien a altres versions que eren bastant diferents entre elles, com per exemple el diari d'Ana Frank, on algunes versions eren exageradament més llargues que altres, i una versió de Harry Potter en anglès, la qual tenia molt més contingut que les altres.
- Mostres per detectar llengües:
    - Expositius: biografia de Lord Byron, pàgina de Wikipedia sobre la Segona Guerra Mundial, Felis silvestris catus (gat) i Declaració Universal dels Drets Humans.
    - Narratius: un fragment del tercer capítol de Revolta dels animals. El primer capítol del llibre *El ninot de neu* de Jo Nesbø. També altres fragments dels textos que utilitzem per al treball

FONTS DE DADES

Per Internet, en els PDF dels llibres que hem buscat i pàgines com Wikipedia pels textos expositius

PROCEDIMENT PER RECOLLIR LES DADES

Hem fet una selecció de quins fragments fem ús per utilitzar com a font d'informació per crear un perfil estadístic de cada llengua. Cal destacar que la quantitat de paraules i caràcters que disposem per la mostra és molt gran, llavors, esperem bons resultats, com a mínim, per la detecció del text d'algun dels llibres que hem fet servir.

Al final, la decisió ha sigut: dos capítols de Harry Potter, dos del Petit Príncep, un del Quixot i dos de la Revolta dels animals.

Després, mitjançant un full de càlcul, hem introduït els textos que hem triat i hem comptat tota la informació de cada text: nombre de caràcters, nombre de paraules, mitjana de caràcters per paraula, nombre de consonants, totals i una a una, nombre de vocals, totals i una a una, i llargada de les paraules.

Un fet que hem de tenir en compte és que vam determinar que una paraula és una quantitat de caràcters indefinits sense cap separació pel mig com un guió o un apòstrof.

Ho hem fet de la següent manera:

- Introduïm textos, els quals s'han substituït totes els apòstrofs, guions i signes de puntuació per espais perquè el recompte de la llargada sigui més senzill.

	Longitud text en curs	
Text en curs	n.caràcters	n.paraules
El señor Jo	14375	2547

aux SPLIT	El	señor	J
aux n.lletres	2	5	

- Nombre de caràcters amb espais: amb el comandament LARGO(A8). Compta tots els caràcters del text.
- Nombre de paraules: amb el comandament CONTARA(F8:8). Compta totes les caselles d'aux SPLIT cap a la dreta.
- Aux SPLIT: amb el comandament SPLIT(A8;" "). Separa per paraules tot el text a les caselles de la dreta.
- Aux n.lletres: Compta la quantitat de lletres de les paraules de les caselles de dalt.

aux SPLIT	El	señor	Jones
aux n.lletres	2	5	5

n. lletres	1	2	3
n. paraules	163	523	407

- Per calcular la quantitat de paraules amb un cert nombre de lletres hem fet servir el comandament: CONTAR.SI(F9:$9;F11). Bàsicament, agafa totes les caselles d'aux n.lletres i compta quantes vegades surt el nombre de lletres de cada cas.

n. lletres	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22		Total
n. paraules	163	523	407	199	311	306	234	214	96	58	18	14	4	0	0	0	0	0	0	0	0	0		Proporc
n.paraules T1	163	523	407	199	311	306	234	214	96	58	18	14	4	0	0	0	0	0	0	0	0	0		2547
n.paraules T2	167	542	559	277	264	314	228	111	112	49	28	23	5	3	2	0	2	0	0	0	0	0		2686
n.paraules T3	86	473	649	565	319	232	170	85	51	35	16	7	1	2	0	1	0	0	0	0	0	0		2692
n.paraules T4	68	615	402	410	333	267	265	151	113	45	38	20	7	5	2	1	1	0	0	0	0	0		2743
n.paraules T5	0	179	703	392	391	338	155	138	85	54	45	24	15	11	1	3	4	1	3	0	0	0		2542
Castellà	6,40%	20,53%	15,98%	7,81%	12,21%	12,01%	9,19%	8,40%	3,77%	2,28%	0,71%	0,55%	0,16%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%		
Català	6,22%	20,18%	20,81%	10,31%	9,83%	11,69%	8,49%	4,13%	4,17%	1,82%	1,04%	0,86%	0,19%	0,11%	0,07%	0,00%	0,07%	0,00%	0,00%	0,00%	0,00%	0,00%		
Anglès	3,19%	17,57%	24,11%	20,09%	11,85%	8,63%	6,22%	3,16%	1,88%	1,20%	0,59%	0,26%	0,04%	0,07%	0,00%	0,04%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%		

- Es recullen totes les dades de cada idioma i després es fa la proporció en base al total.

a	à	á	â	Sum a	e	é	è	ë	ê	Sum e	i	í	ï	î	Sum i	o	ó	ò	ö	ô	Sum o	u	ú	ü	û	Sum u	Vocals
1591	0	0	0	1644	1377	47	0	0	0	1424	536	88	0	0	624	1071	55	0	0	0	1126	475	28	0	0	503	5321
A	À	Á	Â		E	É	È	Ë	Ê		I	Í	Ï	Î		O	Ó	Ò	Ö	Ô		U	Ú	Ü	Û		
9	0	0	0	9	24	0	0	0	0	24	8	0	0	0	8	1	0	0	0	0	1	1	0	0	0	1	43
1600	0	0	0	1653	1401	47	0	0	0	1448	544	88	0	0	632	1072	55	0	0	0	1127	476	28	0	0	504	5364

- Per obtenir les vocals, fem servir el comandament: =LARGO(A7)-LARGO(SUSTITUIR(A7;B6;"")). Agafa la quantitat de la vocal corresponent i després busca el caràcter específic. Per exemple, el comandament posat anteriorment, agafa la vocal A i posteriorment busca específicament l'a minúscula.
- Es fa la suma de totes les vocals: accents, circumflexos, minúscules i majúscules.
    - Un problema que vam tenir és que la lletra a amb accent agut no la vam posar i la vam haver d'afegir posteriorment perquè la vam oblidar, ja que la seva presència només estava en el castellà. Llavors, està col·locada al final del document.

Altres	Consonants	b	c	ç	d	f	g	h	j	k	l	m	n	ñ	p	q	r	s	t	v	w	x	y	z	ß	á
	6308	215	454	0	567	46	119	157	63	0	642	346	780	32	256	120	826	907	487	118	0	14	121	38	0	53
		B	C	Ç	D	F	G	H	J	K	L	M	N	Ñ	P	Q	R	S	T	V	W	X	Y	Z		Á
	156	12	21	0	8	0	1	6	9	0	12	5	16	0	12	3	3	13	20	4	1	0	10	0		0
2548	6464	227	475	0	575	46	120	163	72	0	654	351	796	32	268	123	829	920	507	122	1	14	131	38	0	53

- Per obtenir les consonants, fem servir el comandament: LARGO(A7)-LARGO(SUSTITUIR(A7;AF6;"")). Agafa la quantitat de la consonant i després especifica la que triem, o sigui, minúscula i majúscula.
- Per Altres, fem servir el comandament: LARGO(A7)-AC10-AE10. Compta tots els caràcters i resta les consonants i vocals. Són els espais i altres símbols.
- Les consonants blaves són les exclusives de certs idiomes: el castellà amb la ñ, el català amb la ç i l'alemany amb la ß, encara que, de fet, aquests

caràcters no ens interessen especialment perquè no els utilitzem per detectar la llengua d'un text nou.

- Es fa la suma de totes les consonants: minúscules i majúscules.

Vocals	Altres	Consonants	b	c	ç	d	f	g	h	j	k	l	m	n	ñ	p	q	r	s	t	v	w	x	y	z	ß	á	
5321		6308	215	454	0	567	46	119	157	63	0	642	346	780	32	256	120	826	907	487	118	0	14	121	38	0	53	
			B	C	Ç	D	F	G	H	J	K	L	M	N	Ñ	P	Q	R	S	T	V	W	X	Y	Z		Á	
43		156	12	21	0	8	0	1	6	9	0	12	5	16	0	12	3	3	13	20	4	1	0	10	0		0	
5364	2548	6464	227	475	0	575	46	120	163	72	0	654	351	796	32	268	123	829	920	507	122	1	14	131	38	0	53	
																												SUMA Consonan
5364	2548	6464	227	475	0	575	46	120	163	72	0	654	351	796	32	268	123	829	920	507	122	1	14	131	38	0	53	11828
5100	2822	6740	138	356	27	381	76	161	125	57	1	849	408	795	0	286	177	862	1014	702	252	1	48	19	5	0	0	11840
4218	2711	6984	183	220	0	468	260	239	699	28	76	516	335	763	0	139	9	705	739	978	120	273	13	214	7	0	0	11202
5610	3061	6954	119	393	7	443	113	106	140	70	0	693	370	950	0	305	143	843	980	931	228	1	74	21	24	0	0	12564
4957	2563	7879	220	383	0	673	210	325	632	55	168	466	337	1384	0	66	3	906	802	745	97	239	5	6	132	25	0	12836
45,35%		54,65%	3,51	7,35	0,00	8,90	0,71	1,86	2,52	1,11	0,00	10,1	5,43	12,3	0,50	4,15	1,90	12,8	14,2	7,84	1,89	0,02	0,22	2,03	0,59	0,00	0,99	
43,07%		56,93%	2,05	5,28	0,40	5,65	1,13	2,39	1,85	0,85	0,01	12,6	6,05	11,8	0,00	4,24	2,63	12,7	15,0	10,4	3,74	0,01	0,71	0,28	0,07	0,00	0,00	
37,65%		62,35%	2,62	3,15	0,00	6,70	3,72	3,42	10,0	0,40	1,09	7,39	4,80	10,9	0,00	1,99	0,13	10,0	10,5	14,0	1,72	3,91	0,19	3,06	0,10	0,00	0,00	
44,65%		55,35%	1,71	5,65	0,10	6,37	1,62	1,52	2,01	1,01	0,00	9,97	5,32	13,6	0,00	4,39	2,06	12,1	14,0	13,3	3,28	0,01	1,06	0,30	0,35	0,00	0,00	
38,62%		61,38%	2,79	4,86	0,00	8,54	2,67	4,12	8,02	0,70	2,13	5,91	4,28	17,5	0,00	0,84	0,04	11,5	10,1	9,46	1,23	3,03	0,06	0,08	1,68	0,32	0,00	

Es fa la proporció de vocals i consonants sobre el total, que es troba a la dreta del tot, i cada caràcter específic sobre el total de vocals o consonants.

Consonants i vocals	Mitjana de caràcter		Número de paraules
11828	Castell	4,643894778	2547
11840	Català	4,408041698	2686
11202	Anglès	4,161218425	2692
12564	Francé	4,580386438	2743
12836	Aleman	5,049567270	2542

- Obtenim la mitjana de caràcters per paraula dividint el nombre total de consonants i vocals entre el nombre de paraules.
    - Un càlcul que hem canviat ha sigut la mitjana de caràcters per paraula, perquè ho fèiem dividint el total de caràcters amb espais entre el nombre de paraules. Però vam decidir canviar-ho, perquè vam pensar que els espais no formen part de les paraules i donaria un millor resultat sense espais.

També s'ha de tenir en compte que el català disposa d'un text menys, perquè el Quixot en català no hi és, llavors tenim dades menys fiables en aquest idioma.

Vocals (%)	Animal farm	Harry Potter 1	Harry Potter 2	El Petit Príncep	El Quijote
Castellà	45,35%	45,83%	46,17%	47,04%	45,86%
Català	43,07%	43,07%	43,90%	44,97%	
Anglès	37,65%	36,39%	36,43%	37,90%	38,98%
Francès	44,65%	44,20%	44,42%	45,30%	44,88%
Alemany	38,62%	37,69%	37,28%	38,08%	38,98%

Consonants (%)	Animal farm	Harry Potter 1	Harry Potter 2	El Petit Príncep	El Quijote
Castellà	54,65%	54,17%	53,83%	52,96%	54,14%
Català	56,93%	56,93%	56,10%	55,03%	
Anglès	62,35%	63,61%	63,57%	62,10%	61,02%
Francès	55,35%	55,80%	55,58%	54,70%	55,12%
Alemany	61,38%	62,31%	62,72%	61,92%	61,02%

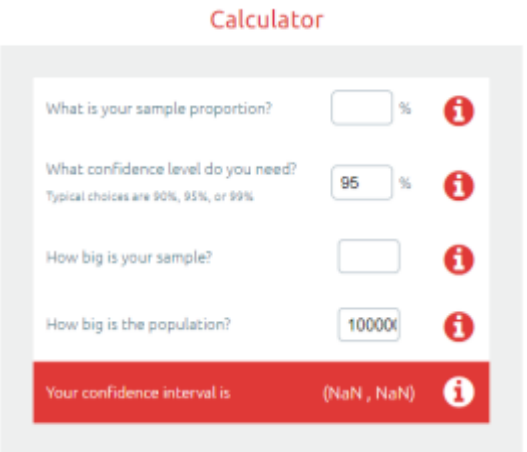
Amb totes les dades, les apuntem en una taula, com per exemple en les taules de dalt amb percentatge de consonants i vocals, i fem la mitjana i calculem els intervals per poder fer, en últim lloc, les gràfiques.

Els intervals de confiança són els números en els quals es mouen uns percentatges o unes dades, perquè així es representa la població a partir d'una mostra. Per fer aquests intervals, utilitzem una calculadora estadística on principalment hem fet servir dos tipus:

Per als que tenim una proporció, o sigui percentatges (variables qualitatives com percentatge de vocals, consonats, etc.), utilitzem la de "population proportion". En aquesta hem de posar el tant per cent que volem calcular, per exemple el percentatge de "a" en un text respecte a la resta de vocals, i després la quantitat de mostra que hem fet servir. Seguint l'exemple de percentatge de *a* respecte les vocals, hauríem d'escriure el total de vocals.

El nivell de confiança el deixem a 95% perquè un nivell de precisió molt alt ens podria deixar fora números necessaris. En l'apartat de població deixem el nombre més gran possible perquè, per exemple, en el cas de les *a,* la població seria la quantitat de vocals en tots els textos escrits en aquell idioma, cosa que és gairebé impossible de calcular.

Pels que tenim la mitjana (variable quantitativa, com el nombre de caràcters), utilitzem la de "population mean". En aquest, al primer apartat hem de posar la mitjana de l'idioma que volem calcular els intervals, després, a la desviació estàndard, s'ha de posar quant varia el nombre de caràcters depenent de l'idioma, o sigui, la dispersió, la qual calculem amb l'aplicació GeoGebra. A la grandària de la mostra, com abans, hem de posar el nombre total de caràcters de tots els textos d'un idioma, cosa que és impossible, així que posem el nombre més gran que podem.



Population proportion - Confidence interval

Population mean - Confidence interval

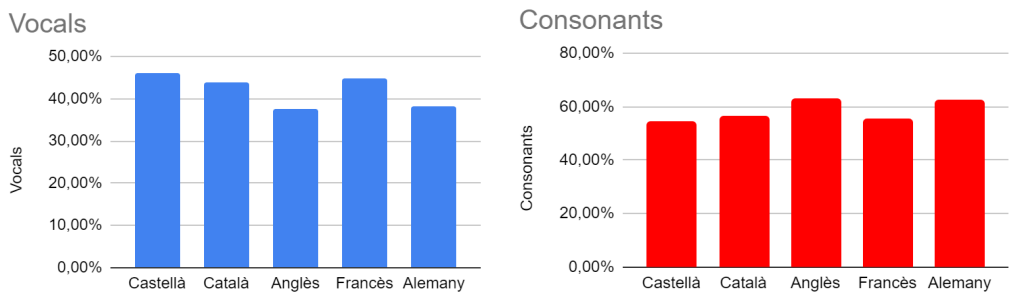
6. ANÀLISI DE DADES

PROCEDIMENT PER ANALITZAR LES DADES

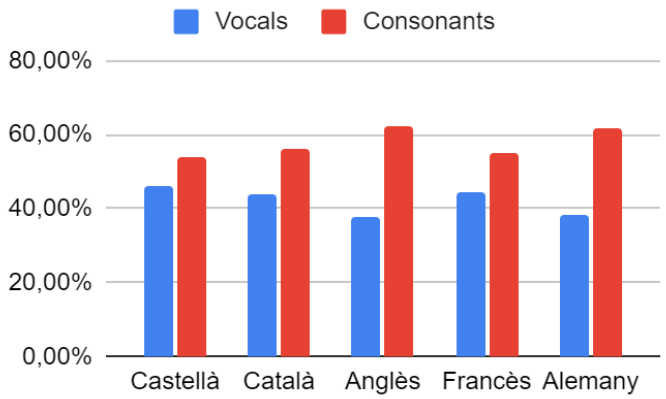
6.1. Creació dels perfils estadístics de les llengües treballades

Buscant característiques específiques per a cada llengua, hem anat creant gràfics per veure quina dada pot sobresortir més en cada aspecte que hem trobat. Bastants dels trets que hem buscat, sobretot els quantitatius, no han servit per determinar un perfil concret per cada llengua perquè els valors no han donat informació rellevant o les dades se superposen i coincideixen, però, per altra banda, tots els altres paràmetres ens donen senyals d'una bona detecció.

➢ Percentatge de vocals i consonants:



➢ Percentatge de vocals i consonants juntes:



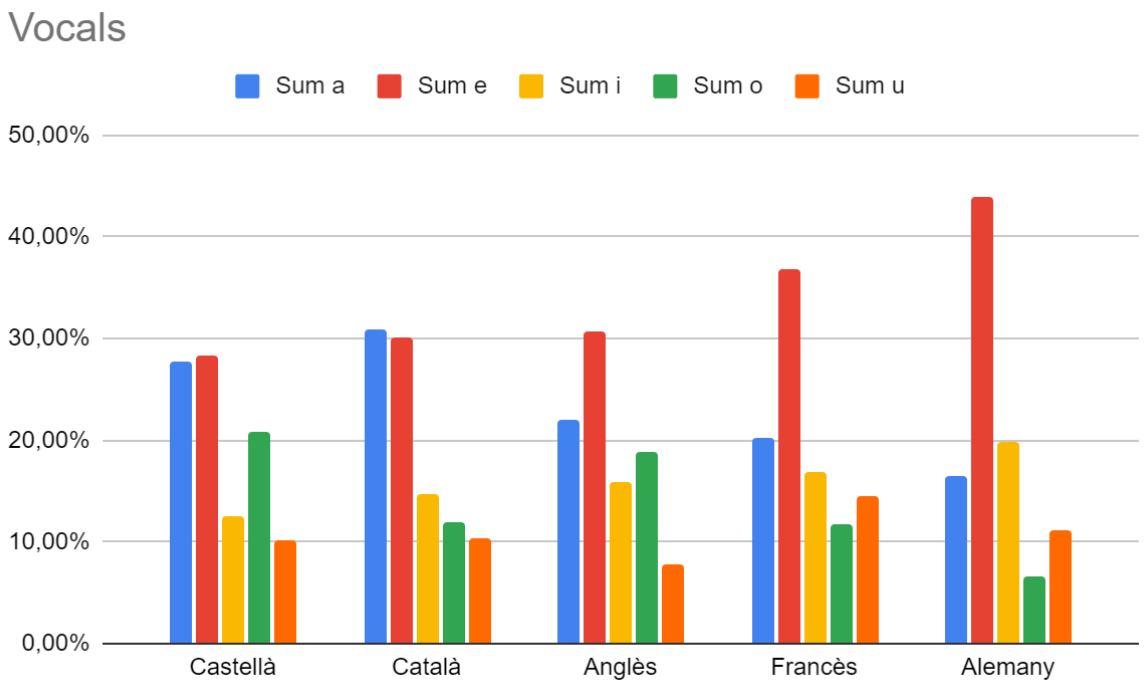
Gràficament, no sobresurt cap idioma ni hi ha cap indici d'una diferència. El màxim que veiem és que el català, el castellà i el francès tenen més vocals que l'alemany i

l'anglès i amb les consonants passa al revés. Això no obstant, els intervals de confiança no se superposen, encara que són molt propers.

	Intervals vocals	Intervals consonants
Castellà	(45.78 , 46.32)	(53.68 , 54.22)
Català	(43.48 , 44.02)	(55.98 , 56.52)
Anglès	(37.21 , 37.73)	(62.27 , 62.79)
Francès	(44.42 , 44.96)	(55.04 , 55.58)
Alemany	(37.87 , 38.39)	(61.61 , 62.13)

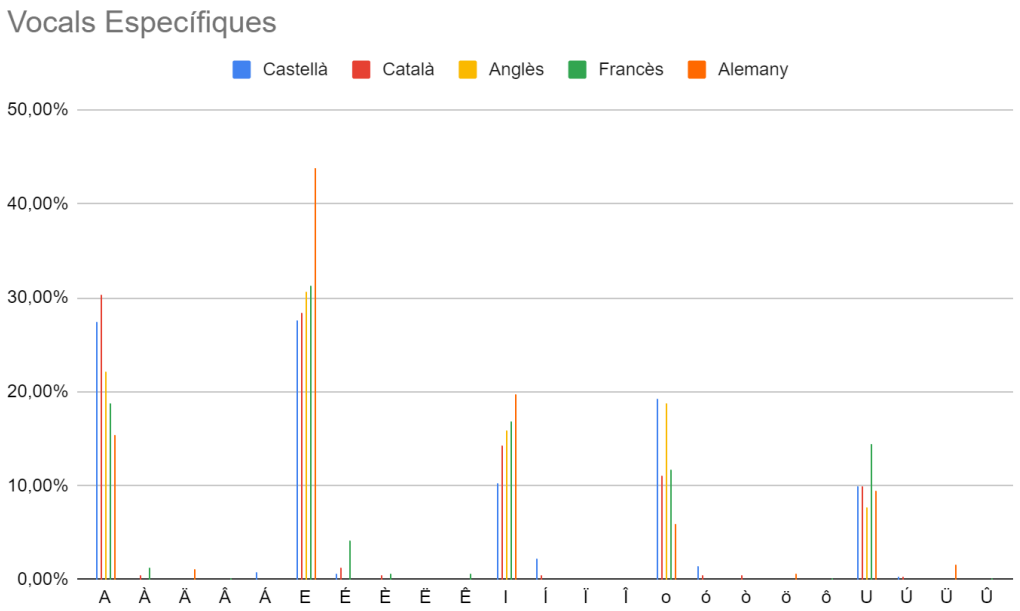
Podem veure una diferenciació de llengües romàniques i germàniques en els percentatges de vocals i consonants.

	Intervals vocals	Intervals consonants	Classificació
Castellà	(45.78 , 46.32)	(53.68 , 54.22)	
Català	(43.48 , 44.02)	(55.98 , 56.52)	
Francès	(44.42 , 44.96)	(55.04 , 55.58)	Romànica
Anglès	(37.21 , 37.73)	(62.27 , 62.79)	
Alemany	(37.87 , 38.39)	(61.61 , 62.13)	Germànica



Ara ens centrem en les vocals específicament, tenim un gràfic on podem veure diferents trets per diferenciar a primera vista:

Llengües	Tret de les vocals a primera vista
Castellà	Té més “O” i menys “I”
Català	Té més “A”
Francès	Té més “U”
Anglès	Té menys “U”
Alemany	Té menys “A” i ”O” i més ”E” i ”I”



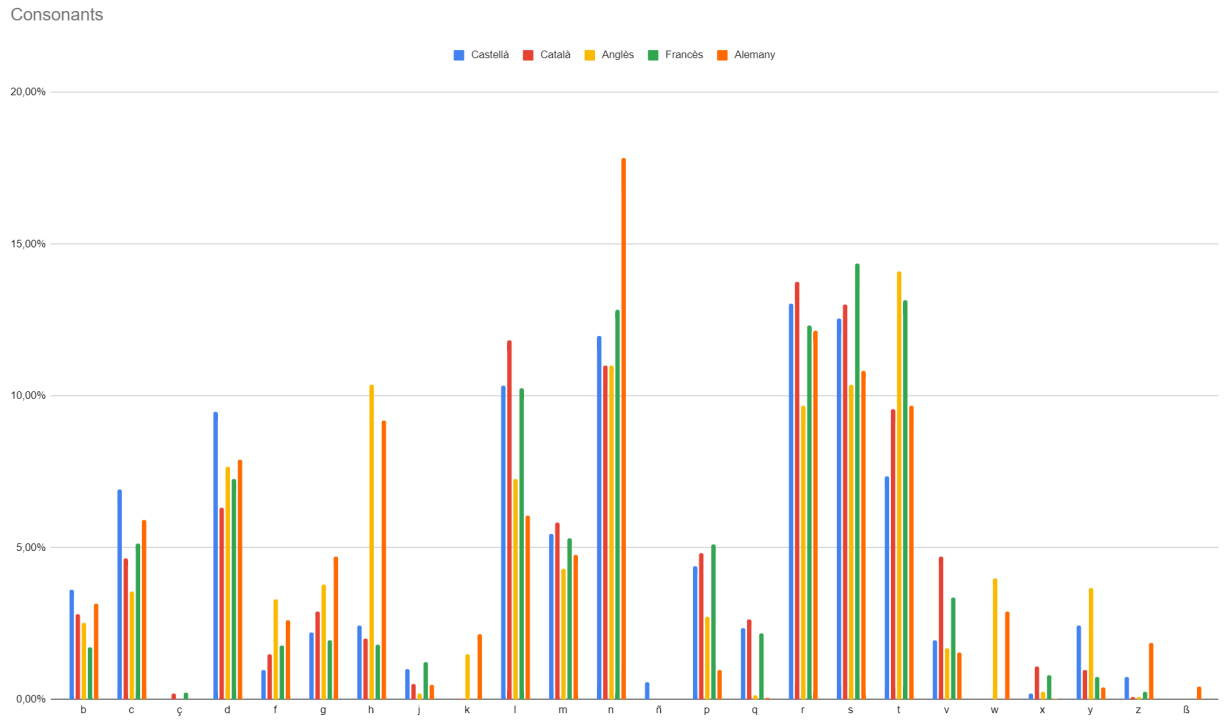
Agafem els intervals de les vocals:

	Intervals Sum a	Intervals Sum e	Intervals Sum i	Intervals Sum o	Intervals Sum u
Castellà	(27.21 , 28.29)	(27.76 , 28.86)	(12.11 , 12.91)	(20.3 , 21.28)	(9.85 , 10.59)
Català	(30.25 , 31.43)	(29.61 , 30.77)	(14.33 , 15.23)	(11.59 , 12.41)	(9.97 , 10.75)
Anglès	(21.5 , 22.66)	(30.04 , 31.34)	(15.4 , 16.42)	(18.26 , 19.36)	(7.34 , 8.08)
Francès	(19.69 , 20.61)	(36.25 , 37.37)	(16.54 , 17.4)	(11.4 , 12.14)	(14.18 , 15)
Alemany	(16.17 , 16.65)	(43.54 , 44.2)	(19.51 , 20.03)	(6.45 , 6.77)	(10.85 , 11.27)

- Els colors subratllats ( ) són els que se superposen.

Gràcies als intervals, tenim nous paràmetres que ens poden indicar l'idioma del text.

➢ Ens fixem en les consonants.



Llengües	Tret de les consonants a primera vista
Castellà	Té més b, c i d.
Català	Té més l, v, q, x i r.
Francès	Té més s i menys b.
Anglès	Té més h, t, f, w i y.
Alemany	Té més n, z, g i k.

Agafem els intervals dels trets de les consonants, que són els valors els quals sobresurten més i llavors ens assegurem que almenys un interval no se solapi amb un altre perquè és més diferent.

	Intervals b	Intervals c	Intervals d	Intervals f	Intervals g	Intervals h
Castellà	(3.4 , 3.82)	(6.63 , 7.19)	(9.13 , 9.79)	(0.86 , 1.08)	(2.06 , 2.38)	(2.26 , 2.6)
Català	(2.64 , 3)	(4.42 , 4.88)	(6.05 , 6.59)	(1.36 , 1.62)	(2.7 , 3.08)	(1.84 , 2.16)
Anglès	(2.35 , 2.69)	(3.36 , 3.76)	(7.38 , 7.94)	(3.12 , 3.5)	(3.59 , 3.99)	(10.03 , 10.69)
Francès	(1.59 , 1.85)	(4.9 , 5.36)	(6.99 , 7.53)	(1.65 , 1.93)	(1.79 , 2.07)	(1.67 , 1.95)
Alemany	(3 , 3.3)	(5.72 , 6.12)	(7.66 , 8.12)	(2.48 , 2.76)	(4.54 , 4.9)	(8.94 , 9.44)

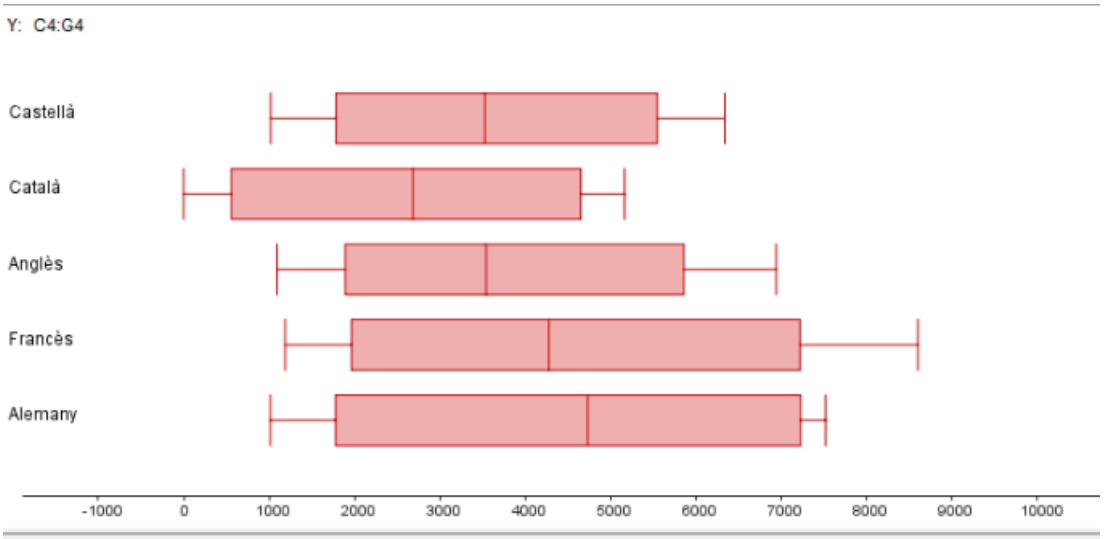
	Intervals j	Intervals k	Intervals l	Intervals n	Intervals q	Intervals r
Castellà	(0.9 , 1.12)	(0 , 0.02)	(10 , 10.68)	(11.62 , 12.34)	(2.19 , 2.53)	(12.66 , 13.42)
Català	(0.44 , 0.6)	(0 , 0.02)	(11.48 , 12.2)	(10.63 , 11.33)	(2.45 , 2.81)	(13.37 , 14.13)
Anglès	(0.15 , 0.25)	(1.37 , 1.63)	(6.99 , 7.55)	(10.66 , 11.32)	(0.09 , 0.17)	(9.36 , 10)
Francès	(1.11 , 1.33)	(0 , 0.02)	(9.93 , 10.55)	(12.48 , 13.18)	(2.04 , 2.34)	(11.98 , 12.66)
Alemany	(0.43 , 0.55)	(2.03 , 2.27)	(5.85 , 6.27)	(17.51 , 18.17)	(0.03 , 0.07)	(11.85 , 12.41)

	Intervals s	Intervals t	Intervals v	Intervals w	Intervals x	Intervals z
Castellà	(12.16 , 12.9)	(7.05 , 7.63)	(1.81 , 2.11)	(0 , 0.02)	(0.15 , 0.25)	(0.65 , 0.85)
Català	(12.64 , 13.38)	(9.24 , 9.9)	(4.48 , 4.96)	(0 , 0.02)	(0.98 , 1.22)	(0.04 , 0.1)
Anglès	(10.02 , 10.68)	(13.74 , 14.48)	(1.54 , 1.82)	(3.79 , 4.21)	(0.21 , 0.31)	(0.05 , 0.11)
Francès	(14 , 14.72)	(12.8 , 13.5)	(3.16 , 3.54)	(0 , 0)	(0.72 , 0.9)	(0.2 , 0.3)
Alemany	(10.54 , 11.08)	(9.42 , 9.92)	(1.42 , 1.64)	(2.76 , 3.04)	(0 , 0.02)	(1.75 , 1.99)

- Els colors subratllats (/)són els que se solapen.

Ens fixem en nombre de paraules, caràcters i mitjana de caràcters per paraula.

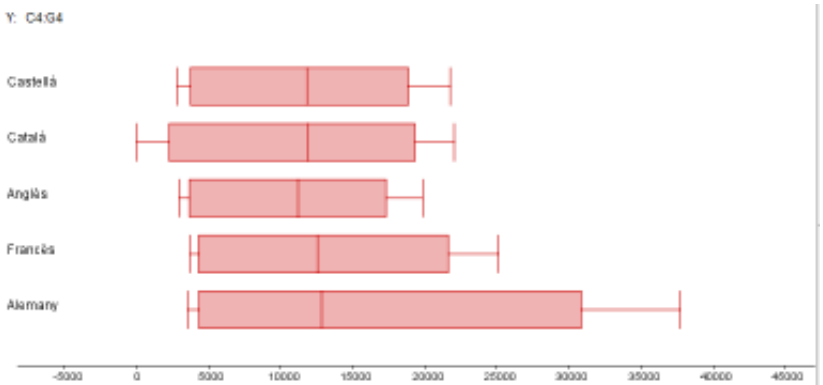
➢ Nombre de paraules:



Estadística inferencial

	n	Mitjana	σ	s	Mín	Qt1	Mediana	Q3	Màx
Castellà	5	3637.2	1821.9947	2037.052	1020	1783.5	3527	5546	6340
Català	5	2622	1892.1454	2115.4829	0	560	2686	4652	5164
Anglès	5	3808.6	1970.619	2203.2191	1095	1893.5	3543	5856.5	6940
Francès	5	4529.8	2556.7604	2858.545	1190	1966.5	4279	7218.5	8600
Alemany	5	4547.6	2491.3175	2785.3776	1016	1779	4731	7224.5	7519

➢ Nombre de caràcters:



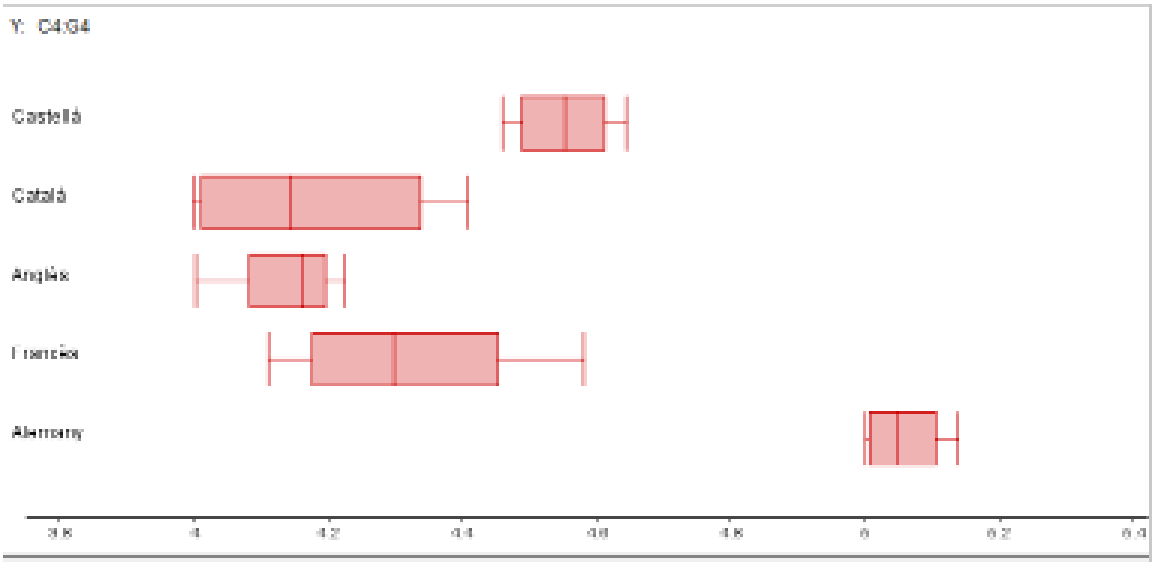
Estadística inferencial

	n	Mitjana	σ	s	Mín	Qt1	Mediana	Q3	Màx
Castellà	5	11398.4	7030.6748	7860.5334	2828	3736.5	11828	18845.5	21760
Català	5	10988.6	7958.6456	8898.0363	0	2251.5	11840	19300	22039
Anglès	5	10628.2	6338.8395	7087.038	2932	3657.5	11202	17312	19861
Francès	5	12883	8057.6816	9008.7618	3721	4306.5	12564	21619	25090
Alemany	5	16637.2	12781.998	14290.7082	3561	4321	12836	30854	37698

Veiem que els valors de nombre de paraules i caràcters no ens serveixen com a criteri per diferenciar la llengua d'un text, perquè els valors no tenen cap sentit com

a indicador d'un idioma o altre, perquè si introduïm un text que tingui un nombre més petit de paraules que el que tenim, no detectarà res i, per tant, és un criteri nul.

➢ Nombre de caràcters/paraula:



Estadística inferencial

	n	Mitjana	σ	s	Min	Qt1	Mediana	Q3	Max
Castellà	5	4.5509	0.0613	0.0685	4.4606	4.4887	4.5539	4.6115	4.6439
Català	4	4.1742	0.1713	0.1978	4.0002	4.0104	4.1442	4.3379	4.408
Anglès	5	4.1433	0.0743	0.083	4.0027	4.0819	4.1612	4.1958	4.2248
Francès	5	4.3115	0.1535	0.1717	4.1109	4.1761	4.2984	4.4536	4.5804
Alemany	5	5.0556	0.0491	0.0548	5.001	5.0073	5.0496	5.1068	5.1385

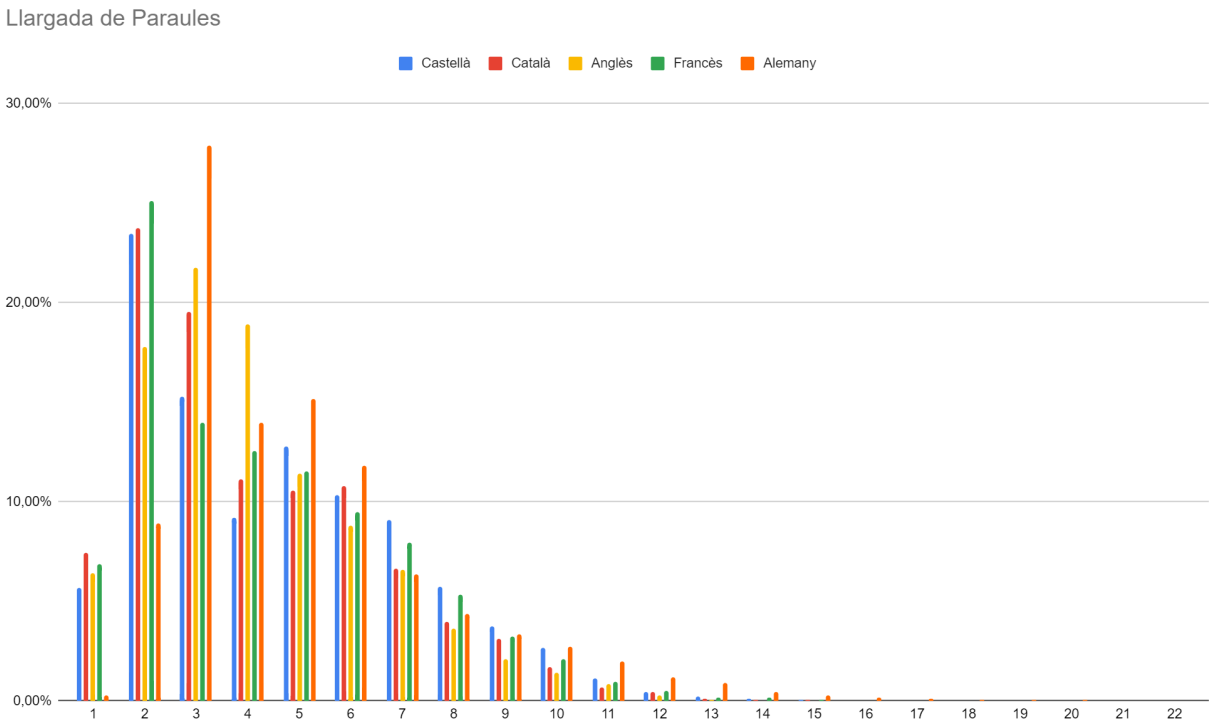
A primera vista, el primer que es pot veure és que l'alemany té molts més caràcters per paraula que la resta d'idiomes. A més, és el que té menys dispersió. L'altre idioma que està una mica més apartat de la resta i té menys dispersió és el castellà. Finalment, l'anglès també té poca dispersió tot i que aquest, com el català i com el francès estan molt junts, o sigui, que se solapen.

Mirem els intervals:

	Intervals
Castellà	(4.53 , 4.58)
Català	(4.09 , 4.26)
Anglès	(4.11 , 4.18)
Francès	(4.25 , 4.38)
Alemany	(5.04 , 5.08)

- Els colors subratllats ( )són els que se solapen.

Fem un cop d'ull a la llargada de les paraules:



Mirem els intervals de confiança:

	Intervals 1	Intervals 2	Intervals 3	Intervals 4	Intervals 5	Intervals 6
Castellà	(5.37 , 5.95)	(22.92 , 23.98)	(14.83 , 15.73)	(8.83 , 9.55)	(12.34 , 13.16)	(9.97 , 10.73)
Català	(7.11 , 7.75)	(23.22 , 24.26)	(19.05 , 20.01)	(10.76 , 11.52)	(10.18 , 10.92)	(10.4 , 11.16)
Anglès	(6.12 , 6.72)	(17.3 , 18.24)	(21.24 , 22.26)	(18.45 , 19.41)	(11 , 11.78)	(8.43 , 9.13)
Francès	(6.61 , 7.19)	(24.63 , 25.61)	(13.58 , 14.36)	(12.18 , 12.94)	(11.17 , 11.89)	(9.15 , 9.81)
Alemany	(0.21 , 0.33)	(8.6 , 9.22)	(27.43 , 28.39)	(13.59 , 14.33)	(14.78 , 15.56)	(11.46 , 12.16)

	Intervals 7	Intervals 8	Intervals 9	Intervals 10	Intervals 11	Intervals 12
Castellà	(8.71 , 9.43)	(5.46 , 6.04)	(3.52 , 4)	(2.49 , 2.89)	(0.99 , 1.25)	(0.39 , 0.55)
Català	(6.37 , 6.97)	(3.75 , 4.23)	(2.92 , 3.34)	(1.53 , 1.85)	(0.56 , 0.76)	(0.38 , 0.54)
Anglès	(6.3 , 6.9)	(3.4 , 3.86)	(1.91 , 2.25)	(1.26 , 1.54)	(0.76 , 0.98)	(0.22 , 0.34)
Francès	(7.63 , 8.25)	(5.09 , 5.61)	(3.01 , 3.41)	(1.95 , 2.27)	(0.86 , 1.08)	(0.45 , 0.61)
Alemany	(6.12 , 6.64)	(4.15 , 4.59)	(3.16 , 3.54)	(2.54 , 2.9)	(1.82 , 2.12)	(1.04 , 1.28)

	Intervals 13	Intervals 14	Intervals 15	Intervals 16	Intervals 17	Intervals 18
Castellà	(0.19 , 0.31)	(0.09 , 0.17)	(0.03 , 0.09)	(0 , 0)	(0 , 0)	(0 , 0)
Català	(0.09 , 0.17)	(0.02 , 0.08)	(0.01 , 0.05)	(0 , 0)	(0 , 0.04)	(0 , 0)
Anglès	(0.03 , 0.09)	(0 , 0.02)	(0 , 0)	(0 , 0.02)	(0 , 0.02)	(0 , 0)
Francès	(0.11 , 0.19)	(0.1 , 0.18)	(0.01 , 0.05)	(0 , 0.02)	(0 , 0.02)	(0 , 0)
Alemany	(0.79 , 0.99)	(0.36 , 0.5)	(0.23 , 0.35)	(0.11 , 0.19)	(0.1 , 0.18)	(0.03 , 0.07)

- Els colors subratllats (/ /)són els que se solapen.

6.2. Identificació de la llengua de textos

Un cop feta tota la recollida de dades i després d'haver descartat alguns paràmetres (els que donaven solapaments per a les llengües) hem d'escollir els paràmetres més útils que ens serviran per identificar la llengua en què està escrit un text, per comparació amb els perfils estadístics de cada llengua que hem anat trobant.

Les nostres hipòtesis inicials tenen en aquest punt del treball poca importància. El que és rellevant és el que hem anat descobrint en el punt 6.1.

1. **Paràmetre protagonista.**

El primer que hem fet servir és la freqüència de vocals totals del text. En veure que els intervals de confiança no se superposaven, vam decidir que la proporció de vocals seria un bon paràmetre.

Per fer més fàcil la detecció, vam agafar directament la proporció de vocals de cada llengua, no tot l'interval de confiança, perquè si comparàvem amb tot l'interval, a l'hora de programar, els resultats que es quedaven fora de l'interval donaven el resultat de "desconegut". Per això era menys complicat i, alhora més eficient, agafar directament la proporció.

Per fer-ho, hem agafat la proporció de vocals que acostuma a haver-hi en cadascun dels cinc idiomes. Llavors, el que fem és calcular la proporció de vocals del text "fantasma", el text del qual volem esbrinar la llengua.

Un cop ho tenim calculat, a les proporcions de cada idioma els restem la proporció de vocals del text fantasma i posem el resultat en valor absolut. El resultat que més s'aproximi a 0 és la llengua a la qual pertany el text fantasma perquè la proporció serà la més semblant. Aquest procediment no dona sempre un resultat vertader, per això hem de fer servir altres paràmetres per assegurar el resultat.

- Exemples: aquí veiem els resultats d'un text en alemany de la Segona Guerra Mundial extret de Wikipedia amb una llargada de 8895 caràcters. El resultat que dona és que el valor més proper al 0 és el que es troba amb l'alemany. Si fem el mateix text amb una llargada de 4858 caràcters, el resultat encara és el correcte.

- Text amb 8895 caràcters en alemany. (**Alemany, s'aproxima més al zero)**

TOTAL CONSONANTS	TOTAL VOCALS	PROPORCIÓ DE VOCALS AL TEXT	CASTELLÀ	CATALÀ	ANGLÈS	FRANCÈS	ALEMANY
4443	2927	39,715	6,334938942	4,034938942	2,245061058	4,974938942	1,585061058

- Text amb 4858 caràcters en alemany.(**Alemany, s'aproxima més al zero**)

TOTAL CONSONANTS	TOTAL VOCALS	PROPORCIÓ DE VOCALS AL TEXT	CASTELLÀ	CATALÀ	ANGLÈS	FRANCÈS	ALEMANY
2444	1590	39,415	6,635027268	4,335027268	1,944972732	5,275027268	1,284972732

Tot i això, quan la mostra és més petita, ho detecta pitjor. El text és d'Ana Frank en anglès i ho ha detectat en alemany.

- Text en anglès de 542 caràcters: **(Alemany, s'aproxima més al zero)**

TOTAL CONSONANTS	TOTAL VOCALS	PROPORCIÓ DE VOCALS AL TEXT	CASTELLÀ	CATALÀ	ANGLÈS	FRANCÈS	ALEMANY
334	208	38,376	7,673616236	5,373616236	0,9063837638	6,313616236	0,2463837638

Un fet sorprenent, però, és que hem posat un microrelat i ho ha detectat bé.

- "Cuando desperté el dinosaurio todavía estaba ahí" **(Castellà, s'aproxima més al zero).**

TOTAL CONSONANTS	TOTAL VOCALS	PROPORCIÓ DE VOCALS AL TEXT	CASTELLÀ	CATALÀ	ANGLÈS	FRANCÈS	ALEMANY
20	22	52,381	6,330952381	8,630952381	14,91095238	7,690952381	14,25095238

També veiem que si introduïm un fragment d'un text que es trobava en els textos que hem utilitzat com a dades per crear els perfils, ho detecta millor. Per exemple, un fragment del tercer capítol de Revolta dels animals en francès te'l detecta bé.

- Text amb 6144 caràcters en francès. **(Francès, s'aproxima més al zero).**

TOTAL CONSONANTS	TOTAL VOCALS	PROPORCIÓ DE VOCALS AL TEXT	CASTELLÀ	CATALÀ	ANGLÈS	FRANCÈS	ALEMANY
3342	2772	45,339	0,7114327772	1,588567223	7,868567223	0,6485672228	7,208567223

Fins i tot amb la Declaració Universal dels Drets Humans, ho detecta bé si s'introdueix tot el text: articles i preàmbuls. Ho hem provat en castellà i francès, i ha donat bé, encara que el text en anglès l'ha confós amb l'alemany. Nosaltres teníem com a hipòtesi inicial que ho faria pitjor.

Aquest paràmetre, el percentatge de vocals, funciona sobretot en textos en castellà i català. El francès el sol confondre amb el català i després l'alemany i l'anglès també els confon entre ells dos, depenent del text que es posi. També veiem que en la majoria de casos perquè ho detecti millor s'ha de posar un text de més o menys 2000 caràcters, però com més gran, millor.

2. **Paràmetres de confirmació.**

Per intentar confirmar-ho, hem fet servir les consonants que més destacaven a cada idioma i que menys se solapaven amb les altres. Aquests paràmetres han donat resultats similars o fins i tot pitjors que els anteriors. De vegades donen bon senyal i de vegades no.

Consonants per a cada idioma:

- ➢ Alemany: percentatge de K i N.
- ➢ Català: percentatge de V i L.
- ➢ Castellà: percentatge de C i D.
- ➢ Francès: percentatge més baix de B i més alt de S.
- ➢ Anglès: percentatge d'H i de T.

2,15	1,5			4,72	3,35			6,91	5,92			1,72	3,61			10,36	9,19		
CONFIRMACIÓ ALEMANY				CONFORMACIÓ CATALÀ				CONFIRMACIÓ CASTELLÀ				CONFORMACIÓ FRANCÉS				CONFORMACIÓ ANGLÈS			
K%	K Alema	K Altres	IDIOMA	V%	V cat	V altres	IDIOMA	C %	C Castellà	C Altres	IDIOMA	B%	B francé	B altres	IDIOMA	H%	H anglès	H altres	IDIOMA
0,26	1,895	1,24457	ALTRES	1,92	2,804291	1,43	ALTRES	5,87	1,035	0,04515	ALTRES	2,55	0,83427	1,05572	ALTRES	1,53	8,82743	7,65743	ALTRES
17,84	12,83			11,84	10,34			9,46	7,89			14,36	13,01			14,11	13,15		
CONFIRMACIÓ ALEMANY				CONFORMACIÓ CATALÀ				CONFIRMACIÓ CASTELLÀ				CONFORMACIÓ FRANCÉS				CONFORMACIÓ ANGLÈS			
N%	N Alema	N Altres	IDIOMA	L%	L cat	L altres	IDIOMA	D %	D Castellà	D Altres	IDIOMA	S%	S francé	S altres	IDIOMA	T%	T anglès	T altres	IDIOMA
14,05	3,791	1,21853	ALTRES	10,22	1,622886	0,12	ALTRES	8,17	1,286	0,28369	ALTRES	13,03	1,33318	0,01681	ALTRES	13,79	0,31689	0,64310	ANGLÈS

Els paràmetres que et detecta són si és l'idioma de confirmació o qualsevol altre segons si compleix el percentatge característic de cada llengua. Segueix el mateix mètode que l'anterior: el nombre més proper de 0 és l'idioma que detecta.

CONFIRMACIÓ CASTELLÀ			
C %	C Castellà	C Altres	IDIOMA
6,65	0,263	0,72697	CASTELL
9,46	7,89		
CONFIRMACIÓ CASTELLÀ			
D %	D Castellà	D Altres	IDIOMA
8,71	0,745	0,82491	CASTELL

Aquí hem posat un fragment del capítol 7 *Revolta dels animals* en castellà i hem pogut tornar a comprovar que el castellà és un idioma que amb aquests paràmetres te'ls detecta molt bé. A més la resta ha donat "altres", per tant ens dona un indici molt clar del fet que aquest text és castellà.

Després de posar el capítol 3 del Petit Príncep en francès, en l'apartat de detecció mitjançant percentatge de vocals i consonats en diu que és francès, però després amb els paràmetres de confirmació tota l'estona diu "altres" i, llavors, segons això no és cap dels idiomes treballats, cosa que és incorrecte.

Percentatge de paraules amb 14 lletres:

- Es pot veure que dona alemany com hauria de ser, però amb valors molt semblants a les altres llengües.

Paraules amb 14 lletres	CASTELLÀ	CATALÀ	ANGLÈS	FRANCÈS	ALEMANY	IDIOMA
1,732501733	1,60250173	1,682501	1,722501	1,59250173	1,30250173	ALEMAN

Aquest paràmetre té un parell de problemes: el primer és que els textos amb un nombre de paraules molt baix gairebé mai dona correctament, ja que quan la mida del text fantasma és molt baixa, els valors poden variar molt i això provoca que no conti bé. El segon problema és que només pot separar l'alemany si el text conté una paraula o més de 14 lletres que això va relacionat amb l'anterior problema perquè si no té una quantitat alta de paraules és poc probable que tingui una de 14 lletres.

Llengües	Paràmetres per la identificació de textos
Castellà	Intervals de vocals i consonants, freqüència de d i c si concorda amb l'interval de vocals i consonants.
Català	Intervals de vocals i consonants, freqüència de v i l si concorda amb l'interval de vocals i consonants.
Anglès	Intervals de vocals i consonants, freqüència de h i t si concorda amb l'interval de vocals i consonants.
Francès	Intervals de vocals i consonants, freqüència de b i s si concorda amb l'interval de vocals i consonants.
Alemany	Intervals de vocals i consonants, freqüència de k i n si concorda amb l'interval de vocals i consonants, paraules amb 14 lletres si el text fantasma és llarg.

7. CONCLUSIONS GENERALS

7.1. Hipòtesis inicials

- L'anglès és l'idioma amb menys caràcters per paraula o amb paraules més curtes.
  No hi ha una diferència prou significativa
- L'alemany és l'idioma amb paraules més llargues.
  Ha quedat verificada
- L'espanyol és l'idioma que més paraules necessita per expressar una idea.
  S'ha descartat treballar en aquesta línia perquè els recomptes de lletres, longituds de paraula… han absorbit tot el temps que hi hem dedicat perquè semblava més productiu de cara a l'objectiu (identificar l'idioma d'un text)
- El català és l'idioma que més vocals té.
  Falsa
- El francès té gran presència de consonants com la s i la x.
  La presència destacada d' s ha quedat verificada, però no la d'x.
- No es detectaran amb precisió textos com la Declaració Universal dels Drets Humans.

Fals si s'utilitza el text íntegre de la Declaració amb el nostre paràmetre protagonista.

7.2. Hipòtesis que s'han elaborat a partir de l'anàlisi dels textos utilitzats per crear els perfils estadístics de les llengües

- **Paràmetre protagonista: La proporció de vocals totals d'un text permet discernir en quina llengua està escrit el text**
  - *No funciona el 100% de les vegades, però un detall que hem pogut observar és que sempre separa bé les llengües Romàniques de les Germàniques.*
  - *Els textos en castellà i català són detectats molt bé en la majoria de casos.*
  - *Com més quantitat de paraules més eficaç és el programa, ja que en tenir una mostra petita si surt una dada atípica afecta molt.*
  - *Amb els textos que funciona millor és amb els narratius, especialment amb textos fantasma de les mateixes obres utilitzades per crear els perfils.*
  - *Amb els textos expositius funciona una mica pitjor.*

- **Paràmetres de confirmació: Consonants singulars que més destaquen a cada idioma**
  - *Han donat resultats similars o pitjors que el paràmetre protagonista.*

7.3. Objectius

- *Hem trobat paràmetres per fer perfils estadístics de les llengües treballades*
- *Hem escrit un programa de detecció de l'idioma d'un text segons alguns dels paràmetres trobats.*
- *Hem detectat l'idioma de textos amb molts encerts*

8. PROPOSTES DE MILLORA

- Creiem que més varietat de textos de mostra per elaborar el perfil estadístic de cada llengua ajudaria a millorar la predicció de l'idioma del text.
- No hem pogut provar en el programa de detecció la totalitat dels paràmetres de caracterització dels idiomes que hem calculat. Pensem que alguns d'ells poden millorar la detecció.

9. BIBLIOGRAFIA I WEBGRAFIA

➢ Calculadora estadística

➢ Declaració Universal dels Drets Humans
➢ Wikipedia - Segunda Guerra Mundial
➢ Wikipedia - Lord Byron
➢ Wikipedia - Felis silvestris catus

➢ Rebelión en la Granja
➢ La rebel·lió dels animals
➢ Animal Farm
➢ La Ferme des Animaux
➢ Farm der Tire

➢ El Principito
➢ El Petit Princep
➢ The Little Prince
➢ Le petite prince
➢ Der Kleine Prinz

➢ El Diario de Ana Frank
➢ Diari d'Ana Frank
➢ Anne Frank - The Diary Of A Young Girl
➢ Le Journal d'Anne Frank
➢ Anne Frank Tagebuch

➢ Ninot de neu
➢ Le bonhomme de neig

➢ Harry Potter y la piedra filosofal
➢ Harry Potter i la pedra filosofal
➢ Harry Potter and the Sorcerer's Stone
➢ Harry Potter à l'école des Sorciers
➢ Harry Potter und der Stein der Weisen

➢ Don Quijote de la Mancha
➢ Don Quichotte de la Manche
➢ Don Quixote of La Mancha
➢ Don Quijote von der Mancha