Author Profiling

Francisco Rangel
Director: Paolo Rosso

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

¿Gender?

¿Personality traits?

¿Emotions?

¿Age?

¿Native language?

Author Profile... Who is who?

Research lines

- ✓ Language use in the different channels of Internet
- ✓ Automatic Identification of Emotions in Facebook comments
- ✓ Language use by gender, theme, emotion and all the combinations
- ✓ PAN 2013 competition on Author Profiling
- ✓ Experiments with Author Profiling dataset - PAN 2013

El uso del lenguaje en los diferentes canales de Internet

Francisco Manuel Rangel Pardo
Autoritas / UPV
http://www.kicorangel.com

Paolo Rosso
UPV
http://www.dsic.upv.es/~prosso/

autoritas
nuevas ideas, nuevas soluciones

VNIVERSITAT POLITÈCNICA VALÈNCIA

comunica2.0
III Congreso universitario sobre redes sociales Campus Gandia

fundéu BBVA

INTRODUCCIÓN

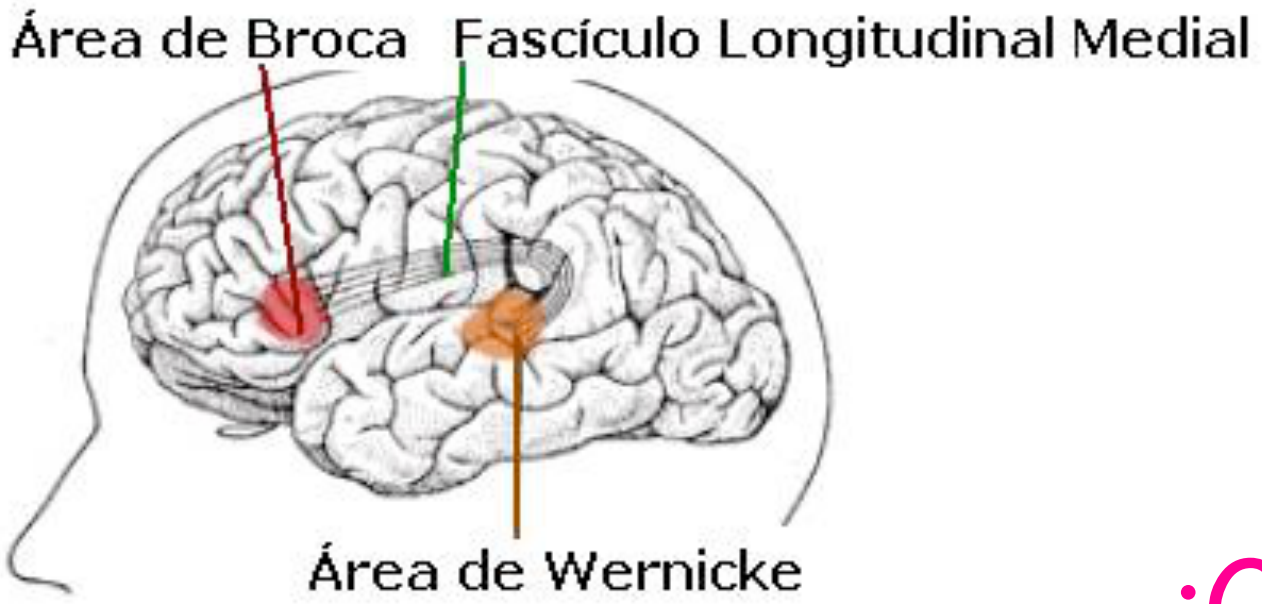
- ✓ El ***estilo discursivo*** es un reflejo de la ***personalidad*** del sujeto que lo elabora. La elección de las palabras y el modo en que se combinan, aporta información de dimensiones como el ***género***, la ***edad*** e incluso el ***estado emocional*** de quién las emite. Pero en Comunicación 2.0 nos enfrentamos a gran variedad de canales y surge la pregunta, ***¿define el canal el modo en que se usa el lenguaje?***

MARCO TEÓRICO

- ✓ **The Secret Life of Pronouns**. James W. Pennebaker
  - ✓ Palabras de contenido 99,96% vs Palabras de función 0,04%
  - ✓ Palabras de función
    - ✓ Cortas y difíciles de detectar
    - ✓ Muy frecuentes
    - ✓ Muy, muy sociales
    - ✓ El cerebro las procesa de manera diferente a las palabras de contenido
- ✓ **Frecuencias del Español. Diccionario y estudios léxicos y morfológicos.** Almela, R., P. Cantos, A. Sánchez, R. Sarmiento, M. Almela
  - ✓ Palabras de contenido 96,92% vs Palabras de función 3.08%
  - ✓ Sustantivos: 54%; Verbos: 22%; Adjetivos: 18%

MARCO TEÓRICO

¿Cómo?



¿Qué?

METODOLOGÍA

- ✓ Se ha determinado un conjunto de ***canales*** de información por sus características representativas de diferentes colectivos de usuarios de Internet: Wikipedia, prensa, blogs, foros, twitter y facebook
- ✓ Se ha recopilado un conjunto significativo de ***documentos*** de dichos canales, todos para el idioma ***Español***
- ✓ Se ha determinado un conjunto de ***categorías gramaticales*** a analizar por su función sintáctica
- ✓ Se ha procesado mediante un POSTagger la extracción automática de la terminología en su correspondiente categoría gramatical
- ✓ Todo lo anterior se ha realizado con la herramienta ***Cosmos*** de ***Autoritas***

METODOLOGÍA: CANALES

DOCUMENTOS	TÉRMINOS
3.987.179	267.465.810

DOCUMENTOS	TÉRMINOS
5.191.694	499.477.658

DOCUMENTOS	TÉRMINOS
1.083.709	122.509.753

DOCUMENTOS	TÉRMINOS
673.664	21.026.388

DOCUMENTOS	TÉRMINOS
23.873.371	163.188.448

DOCUMENTOS	TÉRMINOS
576.723	28.974.716

METODOLOGÍA: CATEGORÍAS GRAMATICALES

- ✓ Adjetivo
- ✓ Adverbio
- ✓ Conjunción
- ✓ Cuantitativo
- ✓ Determinante
- ✓ Interjección
- ✓ Marcador discursivo
- ✓ Preposición
- ✓ Pronombre
- ✓ Sustantivo
- ✓ Verbo

- ✓ Se toma la persona y el número de verbos y pronombres

RESULTADOS: TÉRMINOS ÚNICOS

ÚNICOS	RATIO
162.357	1,89

ÚNICOS	RATIO
157.457	1,83

ÚNICOS	RATIO
162.412	1,89

ÚNICOS	RATIO
93.145	1,08

ÚNICOS	RATIO
128.147	1,49

ÚNICOS	RATIO
110.040	1,28

✓ Lexicon de la RAE: 85.918

RESULTADOS: MEDIA TÉRMINOS POR DOCUMENTO

MEDIA
67

MEDIA
96

MEDIA
113

MEDIA
31

MEDIA
7

MEDIA
50

RESULTADOS: CATEGORÍAS GRAMATICALES

CAT	WIKI	PRENSA	BLOG	FORO	TW	FB
ADJ	13,57%	12,50%	13,67%	**9,27%**	**6,62%**	12,06%
ADV	2,78%	3,46%	3,87%	4,74%	**6,30%**	3,49%
CONJ	1,52%	2,10%	1,80%	4,18%	**7,00%**	2,64%
Q	3,34%	4,47%	4,15%	5,34%	5,53%	4,29%
DET	2,88%	3,48%	2,78%	4,18%	**6,40%**	4,02%
INTJ	0,35%	0,04%	0,06%	0,42%	0,38%	0,07%
MD	0,01%	0,03%	0,02%	0,00%	0,00%	0,00%
PREP	4,00%	5,49%	5,07%	**8,94%**	**13,81%**	6,15%
PRON	0,65%	0,92%	1,12%	2,22%	**3,32%**	1,39%
NOM	**50,33%**	47,05%	46,59%	42,63%	**34,08%**	47,04%
VERB	20,55%	20,47%	20,88%	18,08%	**16,56%**	18,83%

RESULTADOS: PRONOMBRES Y VERBOS

CAT	PER	NUM	WIKI	PRENSA	BLOG	FORO	TW	FB
PRON	1	SIN	13,61%	14,58%	18,85%	**54,47%**	**65,81%**	22,30%
		PLU	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	2	SIN	4,58%	1,18%	2,23%	1,54%	3,53%	3,95%
		PLU	1,92%	1,75%	5,31%	4,61%	5,62%	3,49%
	3	SIN	**55,06%**	**50,75%**	39,26%	24,08%	12,70%	34,68%
		PLU	13,42%	18,22%	16,93%	8,91%	**3,35%**	17,14%
	OTROS		11,41%	13,52%	17,42%	6,39%	8,99%	18,44%
VERB	1	SIN	19,95%	17,41%	17,50%	**28,94%**	**24,00%**	16,61%
		PLU	2,10%	2,42%	4,19%	2,68%	4,68%	4,89%
	2	SIN	6,02%	1,55%	3,58%	3,55%	6,77%	2,95%
		PLU	0,46%	0,42%	0,69%	0,98%	1,65%	0,76%
	3	SIN	31,40%	34,00%	29,92%	28,80%	31,21%	31,21%
		PLU	40,07%	44,20%	45,11%	35,05%	31,69%	43,59%

RESULTADOS: PALABRAS MÁS FRECUENTES

WIKI	PRENSA	BLOG	FORO	TW	FB
de	de	a	de	de	de
en	la	de	y	que	la
la	el	la	que	a	el
y	en	en	a	la	en
el	a	el	la	el	y
por	que	y	el	y	a
un	y	que	en	en	que
una	del	del	un	**no**	los
que	los	los	**no**	**me**	del
a	por	un	**pregunta**	un	por
los	un	por	**es**	**es**	para
del	**se**	**se**	por	**se**	un
es	con	con	**abierta**	**lo**	con
las	las	para	**se**	con	**se**
con	para	las	para	por	**no**

CONCLUSIONES Y TRABAJO FUTURO

- ✓ El presente estudio muestra la variación en el uso del lenguaje según el canal de Internet dónde se comunica
- ✓ El propio canal acentúa el uso de determinadas categorías gramaticales que a su vez son identificativas de claves de personalidad, como el uso de la primera persona, o el uso de las preposiciones
- ✓ Las tablas de distribución deben permitir incorporar un agente correctivo en los estudios de personalidad basados en textos
- ✓ La investigación continúa mediante la división de las categorías en un nivel de detalle mayor, por ejemplo, verbos transitivos, intransitivos, copulativos, auxiliares...
- ✓ El trabajo presentado forma parte de un conjunto de trabajos en identificación de edad y género, extracción de emociones y perfiles de usuario

TRABAJOS RELACIONADOS

- ✓ Competición Author Profiling en PAN2013 (CLEF conference)

  Dado un documento, la tarea consiste en detectar el género y la edad del autor

  http://pan.webis.de

- ✓ Línea de investigación doctoral: Análisis de Emociones y Perfiles de Autor

  Seguir en http://www.kicorangel.com ó @kicorangel

- ✓ Proyecto Cosmos parcialmente financiado por:

ITC/464/2008, TSI-020100-2011-156 y IPT-2012-1220- 430000

Identificación de Emociones en Facebook

Francisco Manuel Rangel Pardo
Autoritas / UPV
http://www.kicorangel.com

Paolo Rosso
UPV
http://www.dsic.upv.es/~prosso/

autoritas®
nuevas ideas, nuevas soluciones

MOTIVACIÓN

- ✓ Conexión entre el estilo de escritura y los rasgos de personalidad (Pennebaker et al. 2003, Oberlander & Gill, 2006)
- ✓ Conexión entre ciertos rasgos de personalidad (pe. psicopatía) y la interpretación y experimentación de ciertas emociones (Hasting et al., 2008, Wilson et al., 2011)

OBJETIVO

- ✓ Determinar un método de identificación automática de emociones en textos en español de un medio social como es Facebook

ESTADO DEL ARTE - RECURSOS

- ✓ Lasswell Value Dictionary (Lasswell & Namenwirth, 1969): Riqueza, poder, rectitud, respeto, iluminación, habilidad, afectividad, bienestar
- ✓ General Inquirer (Stone et al., 1966): Activo, pasivo, fuerte, débil, placer, dolor, sensación, excitación, virtud, vicio, exagerado, subestimado
- ✓ Clairvoyance Affect Lexicon (Huettner & Subasic, 2000): Ira, felicidad, miedo. Centralidad, intensidad
- ✓ Dictionary of Affect in Language (DAL) (Whissell, 1989): 8.742 palabras. Activación e imagen (capacidad de imaginar)
- ✓ Affective Norms for English Words (ANEW) (Bradley & Lang, 1999): Activación, evaluación, control
- ✓ WordNetAffect (Strapparava & Valitutti, 2004): Categoría emocional, evaluación, activación
- ✓ Linguistic Inquiere and Word Count (LIWC) (Pennebaker et. al, 2007): 70 dimensiones. Grado de emociones positivas y negativas, autoreferencia, palabras causales...
- ✓ The Spanish Adaptation of ANEW (Redondo et al., 2007): 1.034 palabras traducidas de ANEW y etiquetadas por 720 participantes
- ✓ Spanish Emotion Lexicon (SEL) (Sidorov et al., 2012): 2.036 palabras con FPA (Probability Factor of Affective use) para cada una de las 6 emociones básicas de Eckman
- ✓ (Osherenko & André, 2007) ¿Los diccionarios afectivos mejoran la identificación o pueden ser sustituidos por diccionarios generales?

ESTADO DEL ARTE: MÉTODOS

- ✓ SEMEval 2007
  - ✓ UPAR7 (Chaumartin, 2007): Parser sintáctico -> tópico principal -> Wordnet
  - ✓ UA (Kozareva et al., 2007): PMI búsquedas web a 3 motores
  - ✓ SWAT (Katz et al., 2007): Unigramas + Roget
- ✓ Comparativa de (Strapparava & Mihalcea, 2008)
  - ✓ WN-AFFECT PRESENCE
  - ✓ LSA SINGLE WORD
  - ✓ LSA EMOTION SYNSET
  - ✓ LSA ALL EMOTION WORDS
  - ✓ NB TRAINED ON BLOGS

	Fine		Coarse	
	r	Prec.	Rec.	F1
WN-AFFECT PRESENCE	9.54	**38.28**	1.54	4.00
LSA SINGLE WORD	12.36	9.88	66.72	16.37
LSA EMOTION SYNSET	12.50	9.20	77.71	13.38
LSA ALL EMOTION WORDS	9.06	9.77	**90.22**	**17.57**
NB TRAINED ON BLOGS	10.81	12.04	18.01	13.22
SWAT	25.41	19.46	8.61	11.57
UA	14.15	17.94	11.26	9.51
UPAR7	**28.38**	27.60	5.68	8.71

ESTADO DEL ARTE: MÉTODOS

- ✓ (Elliot, 1992): Basado en palabras clave
- ✓ (Pang et al., 2002): Afinidad léxica de las palabras a las emociones
- ✓ (Liu et al., 2002): Base de conocimiento OMCS
- ✓ (Dhaliwal et al., 2007): Oraciones imperativas, marcas de exclamación, mayúsculas, presente y futuro
- ✓ (García & Alias, 2008): Arquitectura independiente del idioma basada en módulos
- ✓ (Sugimoto & Yoneyema, 2006): Nombres, adjetivos y verbos en japonés
- ✓ (Mohammad & Yang, 2001): Análisis de emociones por género en cartas de amor, emails de odio y notas de suicidio
- ✓ (Díaz, 2013): Etiquetado basado en diccionario SEL en español y sobre cuentos cortos

ESTADO DEL ARTE: CONCLUSIONES

- ✓ Trabajos centrados principalmente en el inglés
- ✓ Estudios realizados en su mayoría sobre textos tradicionales: prensa, cuentos...
- ✓ Aproximaciones generalmente basadas en características semánticas muy ligadas al contexto o tema, o el uso de diccionarios

COLECCIÓN: RECUPERACIÓN

- ✓ Comentarios de Facebook en Español
- ✓ Recuperación de comentarios de tres temáticas de actualidad y gran activismo:
  - ✓ Política (PSOE, PP, IU, UPyD)
  - ✓ Fútbol (Real Madrid, Barça, Valencia, Athletic de Bilbao)
  - ✓ Personajes públicos (Belén Esteban, Kiko Hernández, David Bisbal y Santiago Segura)

COLECCIÓN: RECUPERACIÓN: DOCUMENTOS

TEMA	PÁGINAS	POSTS	COMENTARIOS
POLÍTICA	PSOE	1.000	22.096
	PP	1.000	4.590
	IU	1.002	2.867
	UPyD ESTUDIANTES	593	135
FÚTBOL	REAL MADRID	1.125	1.035
	BARCELONA FC	1.002	1.520
	VALENCIA FB	1.003	463
	ATHLETIC BILBAO	560	444
PERSONAJES	BELÉN ESTEBAN	1.000	12.191
	SANTIAGO SEGURA	1.000	99
	DAVID BISBAL	1.000	2.902
	KIKO HERNÁNDEZ	1.007	291

COLECCIÓN: ETIQUETADO GÉNERO

- ✓ Con el fin de poder enlazar el estudio de las emociones con el género de los autores, se ha obtenido el género para todos los autores que lo tienen público en su muro, y se procederá a seleccionar la muestra de trabajo a partir de aquellos que lo tengan informado, balanceando entre hombres y mujeres

COLECCIÓN: SELECCIÓN DE MUESTRA

- ✓ Se balancea el mismo número de documentos para cada temática y dentro de cada temática para cada género
- ✓ No se procesan los documentos para no sesgar su selección, sólo se obtienen aquellos que contienen texto en español, aunque se realiza de manera automática y es posible que se cuelen comentarios en algún otro idioma o sólo con urls

TEMA	GÉNERO	COMENTARIOS
POLÍTICA	HOMBRES	200
	MUJERES	200
FÚTBOL	HOMBRES	200
	MUJERES	200
PERSONAJES	HOMBRES	200
	MUJERES	200
TOTAL		1.200

ETIQUETADO: FORMATO

- ✓ Se abre una hoja de excel por cada etiquetador
- ✓ En la primera celda se inserta el comentario y se abre una celda por cada una de las emociones básicas
- ✓ El etiquetador puede seleccionar tantas emociones como considere oportunas
- ✓ Si no considera ninguna emoción, deja en blanco la elección
- ✓ Ejemplo:

comentario	alegría	enfado	miedo	repulsión	sorpresa	tristeza
no tienen vergüenza. Roban y roban a manos llenas y aun encima, se hacen las víctimas.		X		X		

ETIQUETADO: REGLAS

- ✓ Tres formas de entender las emociones en textos:
  - ✓ La emoción con que debería leerse
  - ✓ Qué hace sentir
  - ✓ ***La emoción que se describe o manifiesta en la oración***
- ✓ Para evitar la implicación emocional del etiquetador y la subjetividad dentro de lo posible, se debe realizar el etiquetado intentando seguir el tercer modo de entender las reglas.

ETIQUETADO: REGLAS

- ✓ A continuación las emociones básicas y las secundarias más cercanas para ayudar al etiquetado:

ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA
Agradecido	Agresivo	Acomplejado	Aborrecimiento	Extrañeza	Abatido
Alegre	Colérico	Alarmado	Desagrado	Sobresalto	Agobiado
Animado	Crispado	Angustiado	Grima	Susto	Apenado
Calmado	Descontento	Ansioso	Repulsión	Consternación	Confuso
Confiado	Enfadado	Atemorizado	Antipatía	Pasmo	Decepcionado
Contento	Enojado	Aterrado	Aversión	Desconcierto	Deprimido
Dichoso	Excitado	Avergonzado	Repugnancia	Estupor	Desalentado
Encantado	Fastidiado	Confuso	Disgusto	Asombro	Desanimado
Entusiasmado	Furioso	Desesperado	Repudia	Fascinación	Desdichado
Eufórica	Insatisfecho	Desorientado	Repulsa	Admiración	Desmoralizado
Esperanzado	irascible	Horrorizado	Odio	Confusión	Frustrado
Feliz	Malhumorado	Inquieto	Manía	Chasco	Nostálgico
Gozoso	Molesto	inseguro	Rabia	Impresión	Soledad
Satisfecho	Nervioso	Intranquilo	Animadversión	Exclamación	Triste
Tranquilo	Rabioso	Pánico	Nauseabundo	Conmoción	Infeliz
Complacido	Tenso	Preocupado	Indignación	Estupefacción	Desconsolado
Libre	Violento	Temeroso	Enfado		Afligido
Fascinado	Irritado	Tenso	Desprecio		Amargado
Seguro	Indignado	Indeciso	Distanciamiento		Impotente
		Impotencia			

ETIQUETADO: EVALUACIÓN

- ✓ Dependiendo del número de etiquetadores y de la cantidad de emociones que se pueden etiquetar por texto, tenemos:
  - ✓ Kappa de Kohen: Dos etiquetadores, una emoción por texto
  - ✓ Kappa de Fleiss: Múltiples etiquetadores, una emoción por texto
  - ✓ Kappa de Kohen ponderada: Dos etiquetadores, múltiples emociones por texto
  - ✓ **Kappa de Díaz-Sidorov: Múltiples etiquetadores, múltiples emociones por texto**

CONCORDANCIA

BÁSICAS	ANOTADOR 1	ANOTADOR 2	ANOTADOR 3	RESTO
ANOTADOR 1		0,0587	0,2738	0,1662
ANOTADOR 2	0,0587		0,1042	0,0814
ANOTADOR 3	0,2738	0,1042		0,1890
TOTAL	0,1455			

COMBINADAS	ANOTADOR 1	ANOTADOR 2	ANOTADOR 3	RESTO
ANOTADOR 1		0,6618	0,5656	0,6137
ANOTADOR 2	0,6618		0,5773	0,6196
ANOTADOR 3	0,5656	0,5773		0,5715
TOTAL	0,6016			

ETIQUETADO: SELECCIÓN

- ¿Qué emoción se selecciona para cada comentario?
  - La que haya sido marcada por cualquier anotador (1/3)
  - **La que haya sido marcada por la mayoría de anotadores (2/3)**
  - La que haya sido marcada por todos los anotadores (3/3)

RESULTADOS DEL ETIQUETADO (2/3)

BÁSICAS	ALEGRÍA	ENFADO	MIEDO	REPULSIÓN	SORPRESA	TRISTEZA	NEUTRA
TOTAL	338	151	3	129	390	76	262
%	28,17	12,58	0,25	10,75	32,50	6,33	21,83

COMBINADAS	ALEGRÍA / SORPRESA	ENFADO / REPULSIÓN	MIEDO	TRISTEZA	NEUTRA
TOTAL	639	243	3	76	262
%	53,25	20,25	0,25	6,33	21,83

MODELO APRENDIZAJE: CARACTERÍSTICAS

(F)recuencias: Ratio frente al total del número de palabras únicas, palabras que empiezan en mayúsculas, palabras completamente en mayúsculas, longitud de las palabras, caracteres en mayúsculas y palabras alargadas (pe. Holaaa).
(P)untuación: Frecuencia de uso de puntos, comas, puntos y comas, dos puntos, exclamaciones, interrogaciones y número de comillas.
(C)ategorías gramaticales utilizadas (PoS Tagger). Número y persona de verbos y pronombres, modos verbales, nombres propios (NER) y palabras no identificadas.
(E)moticonos, que aún definiendo el estilo de escritura, incorpora semántica de contenido emocional (Martínez et al., 2012) pero manteniendo independencia del contexto temático (Read, 2005). Ratio del número de emoticonos frente al total de palabras, número de emoticonos alegres, enfadados, disgustados, sorprendidos, triste, burla y mudos[6]

(SEL) Spanish Emotion Lexicon (Sidorov et al., 2012): Para cada palabra del comentario se obtiene su lema y su FPA en el diccionario. Si el lema no tiene entrada, se obtienen sus sinónimos y se obtiene el FPA de los mismos. Se suman todos los FPA para cada emoción.
(BoW): Se obtiene el lema de las 20 primeras palabras con mayor ganancia de información y tratando que sean lo más independientes de la temática posible, para lo que se han obtenido principalmente adjetivos y adverbios, eliminando de estas palabras los sustantivos (pe. gol, recorte...). Se han reducido a su raíz las palabras alargadas (pe. Holaaa) y se ha reducido cualquier combinación de risas (ja, je, ji, jo, ju) y todos sus alargamientos (jajaaj, jejejeje) a una característica común: ja.

ENTRENAMIENTO Y EVALUACIÓN

- ✓ Cuatro algoritmos de aprendizaje
  - ✓ Support Vector Machines
  - ✓ Decission trees
  - ✓ Naïve Bayes
  - ✓ Bayes Net
- ✓ Dos evaluaciones como en SEMEVal 2007
  - ✓ Precision, Recall and F
  - ✓ *r* Kappa de Pearson para correlación

EXPERIMENTOS

- ✓ Experimento 1: Uso de diccionario afectivo
- ✓ Experimento 2: Identificación basada en estilo
- ✓ Experimento 3: Emociones básicas vs. combinadas
- ✓ Experimento 4: Identificación de género

#1: Uso de SEL

		rr	Prec	Rec	F1
Ale.	FPEC	21,2	39,2	**64,5**	48,8
	+SEL	**23,6**	**70,7**	62,9	**64,8**
Enf.	FPEC	**22.98**	**84,8**	72,9	77,0
	+SEL	22,36	84,4	**73,3**	**77,2**
Rep.	FPEC	23,18	**87,1**	74,3	78,7
	+SEL	**23,88**	87,0	**75,6**	**79,6**
Sor.	FPEC	17,8	42,3	56,7	48,4
	+SEL	**18,0**	**64,7**	**60,2**	**61,4**
Tri.	FPEC	15,9	90,9	80,4	84,7
	+SEL	**16,0**	90,9	**80,5**	84,7

- ✓ El uso de diccionario mejora significativamente la identificación en la mayoría de los casos

#2: Identificación basada en estilo

		rr	Prec	Rec	F1
Ale.	J48	27,1	70,6	71,7	71,0
	NB	**27,9**	71,1	68,6	69,5
	BN	25,6	**72,5**	62,6	64,4
	SVM	24,9	71,3	**73,9**	**71,1**
Enf.	J48	16,6	81,9	**84,7**	**83,1**
	NB	**22,6**	**84,5**	73,3	77,2
	BN	22,2	**84,5**	72,8	76,8
	SVM	10,8	80,6	83,8	82,0
Rep.	J48	21,7	85,2	**87,3**	**86,1**
	NB	15,7	85,3	70,8	75,9
	BN	**24,9**	**87,3**	75,9	79,9
	SVM	6,2	81,0	85,4	83,0
Sor.	J48	**25,8**	**67,5**	**67,8**	**67,6**
	NB	20,6	66,6	60,0	61,3
	BN	20,7	66,3	60,7	61,9
	SVM	17,2	64,5	67,3	64,8
Tri.	J48	12,1	89,6	90,9	90,2
	NB	6,1	89,3	75,3	81,1
	BN	**16,7**	**91,1**	80,2	84,5
	SVM	8,2	89,2	**91,6**	**90,3**

- ✓ Las características propuestas, basadas en estilo y de independencia de la temática, proporcionan unos resultados competitivos con el estado del arte (SEMEval) para medios sociales en Español

	Fine	Coarse		
	r	Prec.	Rec.	F1
WN-AFFECT PRESENCE	9.54	**38.28**	1.54	4.00
LSA SINGLE WORD	12.36	9.88	66.72	16.37
LSA EMOTION SYNSET	12.50	9.20	77.71	13.38
LSA ALL EMOTION WORDS	9.06	9.77	**90.22**	**17.57**
NB TRAINED ON BLOGS	10.81	12.04	18.01	13.22
SWAT	25.41	19.46	8.61	11.57
UA	14.15	17.94	11.26	9.51
UPAR7	**28.38**	27.60	5.68	8.71

#3: Emociones básicas vs. combinadas

		π	Prec	Rec	F1
Ale. + Sor.	J48	38,8	69,5	69,6	69,5
	NB	42,1	71,3	71,1	71,1
	BN	40,1	70,5	70,4	70,2
	SVM	**44,5**	**72,9**	**72,1**	**72,1**
Enf. + Dis.	J48	26,0	76,2	**77,3**	**76,7**
	NB	**33,9**	**80,2**	73,0	75,2
	BN	33,0	79,3	73,8	75,7
	SVM	18,9	74,2	**77,3**	75,3

- ✓ Es latente la dificultad de identificación unívoca de emociones cercanas

#4: Identificación de género

	Acc	*r*	Prec	Rec	F1
Hombre	53,6	7,67	54,2	49,7	51,8
Mujer			53,5	53,8	53,8

- ✓ El método propuesto sirve para la identificación del género del autor más allá del puro azar (r=7,67), aunque habrá que esperar para compararlo con resultados como los del PAN

Conclusiones y trabajo futuro

- ✓ Hemos generado un corpus de un medio social como son los comentarios de Facebook y en español, etiquetado en las seis emociones de Eckman
- ✓ Hemos propuesto un método de identificación de emociones basado en características de estilo e independientes de la temática
- ✓ Hemos comprobado la dificultad de identificación de emociones cercanas, como disgusto/enfado y alegría/sorpresa

- ✓ En futuros trabajos verificaremos la propuesta con los datasets de estudios similares (Díaz) y con la competición de Author Profiling de PAN 2013
- ✓ Vamos a indagar más en profundidad en características de estilo de escritura

Language use by gender, theme, emotion and all the combinations (still working on...)

POR GÉNERO

CAT	COMUNICA 2.0	TODO	HOMBRE	MUJER
ADJ	12,06	6,49	6,53	6,45
ADV	3,49	3,93	3,94	3,91
CONJ	2,64	9,51	9,55	9,46
Q	4,29	5,46	5,76	5,12
DET	4,02	7,25	6,81	**7,74**
INTJ	0,07	0,23	0,18	**0,30**
MD	0	0,00	0,00	0,00
PREP	6,15	6,06	**6,25**	5,85
PRON	1,39	2,45	2,24	**2,67**
NOM	47,04	31,89	32,21	31,53
VERB	18,83	15,38	15,44	15,32

POR EMOCIÓN

CAT	TODO	JOY	SURPRISE	FEAR	SADNESS	ANGER	DISGUST
ADJ	6,49	8,43	8,75	9,17	5,07	5,76	5,59
ADV	3,93	3,33	3,41	3,67	4,72	3,80	4,54
CONJ	9,51	7,63	7,51	11,01	11,01	10,66	10,72
Q	5,46	5,71	5,94	**2,75**	5,46	5,35	5,43
DET	7,25	8,01	6,95	8,26	7,19	7,51	7,47
INTJ	0,23	0,65	0,38	0,00	0,13	0,11	0,12
MD	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PREP	6,06	5,86	6,50	4,59	6,02	5,94	6,02
PRON	2,45	3,53	2,73	1,83	2,73	2,53	1,88
NOM	31,89	31,77	31,93	33,03	31,41	32,01	30,74
VERB	15,38	15,33	14,13	17,43	16,68	15,23	15,61

POR TEMA

CAT	TODO	POLITICS	PEOPLE	FOOTBALL
ADJ	6,49	5,74	6,88	**8,89**
ADV	3,93	4,06	3,92	3,37
CONJ	9,51	**10,45**	9,02	6,47
Q	5,46	5,20	**6,78**	3,85
DET	7,25	7,14	7,37	7,46
INTJ	0,23	0,12	**0,43**	0,32
MD	0,00	0,00	0,00	0,00
PREP	6,06	**6,14**	5,92	5,99
PRON	2,45	2,27	**2,86**	2,34
NOM	31,89	32,17	31,99	30,49
VERB	15,38	15,43	**14,90**	16,16

POR EMOCIÓN Y GÉNERO

CAT	JOY		SURPRISE		FEAR		SADNESS		ANGER		DISGUST	
	H	M	H	M	H	M	H	M	H	M	H	M
ADJ	8,28	8,53	8,55	8,91	5,56	10,96	4,95	**5,23**	**6,45**	5,05	**6,28**	4,76
ADV	3,29	3,36	3,40	3,43	5,56	2,74	3,98	5,74	3,76	**3,85**	**4,91**	4,09
CONJ	6,79	**8,15**	7,30	7,68	8,33	12,33	11,33	10,56	10,30	**11,03**	**11,06**	10,31
Q	4,99	**6,16**	5,94	5,94	5,56	1,37	5,48	5,44	**5,64**	5,05	**6,02**	4,71
DET	6,59	**8,90**	6,34	**7,45**	8,33	8,22	7,20	7,18	7,12	**7,92**	7,01	**8,03**
INTJ	0,20	**0,93**	0,40	0,37	0,00	0,00	0,15	0,10	0,18	0,05	0,22	0,00
MD	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PREP	**6,89**	5,23	**7,41**	5,76	2,78	5,48	**6,30**	5,64	5,42	6,49	6,07	5,96
PRON	3,59	3,48	2,72	2,74	0,00	2,74	2,85	2,56	2,24	2,83	1,81	1,97
NOM	33,23	30,86	33,39	30,76	38,89	30,14	31,13	31,79	31,87	32,16	29,13	32,68
VERB	15,27	15,37	13,81	14,40	16,67	16,44	16,80	16,51	15,00	**15,48**	**16,52**	14,50

POR TEMA Y GÉNERO

CAT	POLITICS		PEOPLE		FOOTBALL	
	H	M	H	M	H	M
ADJ	5,84	5,63	6,93	6,82	8,79	8,99
ADV	3,98	4,15	4,11	3,71	3,39	3,36
CONJ	10,59	10,30	8,87	9,18	6,29	6,65
Q	5,68	4,67	6,78	6,78	3,87	3,83
DET	6,98	7,32	6,53	**8,34**	6,61	**8,29**
INTJ	0,11	0,14	0,33	**0,55**	0,16	0,47
MD	0,00	0,00	0,00	0,00	0,00	0,00
PREP	6,00	6,31	6,27	5,52	7,34	4,69
PRON	2,15	2,41	2,49	3,28	2,10	2,58
NOM	32,23	32,11	32,42	31,49	31,69	29,32
VERB	15,46	15,40	14,70	15,12	16,94	15,40

POLÍTICA POR EMOCIÓN Y GÉNERO

CAT	JOY		SURPRISE		FEAR		SADNESS		ANGER		DISGUST	
	H	M	H	M	H	M	H	M	H	M	H	M
ADJ	5,48	8,42	6,81	7,81	0,00	10,96	4,81	4,69	6,20	4,96	7,03	4,69
ADV	4,11	4,40	3,87	4,24	0,00	2,74	3,52	5,12	4,01	3,94	4,43	3,92
CONJ	7,67	10,26	9,02	9,17	5,26	12,33	12,30	10,54	10,87	11,11	10,92	10,22
Q	3,84	6,59	5,89	5,77	10,53	1,37	5,74	4,69	5,27	4,80	5,42	4,63
DET	7,12	5,13	6,45	5,94	10,53	8,22	6,85	7,03	7,19	8,25	7,10	7,78
INTJ	0,27	1,10	0,37	0,34	0,00	0,00	0,09	0,15	0,16	0,05	0,31	0,00
MD	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PREP	6,30	6,23	6,63	8,15	5,26	5,48	6,38	5,12	5,76	6,52	5,96	5,98
PRON	2,47	3,66	2,95	2,04	0,00	2,74	2,87	3,07	2,20	2,59	2,06	2,12
NOM	33,15	29,30	32,23	30,73	52,63	30,14	32,19	33,09	31,50	32,08	29,26	33,03
VERB	16,99	15,38	15,10	15,79	5,26	16,44	15,82	16,11	15,09	15,58	15,97	14,52

PEOPLE POR EMOCIÓN Y GÉNERO

CAT	JOY		SURPRISE		FEAR		SADNESS		ANGER		DISGUST	
	H	M	H	M	H	M	H	M	H	M	H	M
ADJ	10,43	8,51	8,58	9,03	-	-	5,26	6,83	6,49	4,44	5,50	4,97
ADV	4,06	3,06	2,99	3,28	-	-	4,68	6,83	3,25	3,33	5,60	4,97
CONJ	7,25	7,95	7,21	8,32	-	-	7,02	8,84	8,44	11,11	11,52	10,50
Q	6,38	6,58	6,97	7,97	-	-	3,51	7,63	7,79	7,04	7,19	5,25
DET	8,12	9,99	6,22	8,44	-	-	9,36	8,43	6,49	5,19	6,34	9,12
INTJ	0,29	1,14	0,50	0,59	-	-	0,58	0,00	0,32	0,00	0,11	0,00
MD	0,00	0,00	0,00	0,00	-	-	0,00	0,00	0,00	0,00	0,00	0,00
PREP	6,09	5,11	7,21	4,92	-	-	4,68	6,43	2,60	6,67	6,13	5,80
PRON	5,51	3,18	3,61	3,63	-	-	1,75	1,61	2,27	4,81	1,27	1,38
NOM	31,30	31,21	35,95	31,54	-	-	29,24	28,51	34,74	31,85	28,65	30,94
VERB	12,17	14,64	12,81	13,60	-	-	21,05	17,67	13,31	15,56	17,34	14,36

FOOTBALL POR EMOCIÓN Y GÉNERO

CAT	JOY		SURPRISE		FEAR		SADNESS		ANGER		DISGUST	
	H	M	H	M	H	M	H	M	H	M	H	M
ADJ	9,25	8,61	10,71	9,65	11,76	-	6,17	4,65	10,58	15,15	2,90	7,69
ADV	1,37	3,31	3,57	2,95	11,76	-	8,64	9,30	0,96	3,03	4,35	0,00
CONJ	5,14	7,28	5,24	5,76	11,76	-	7,41	20,93	5,77	6,06	7,25	15,38
Q	4,79	5,08	4,05	3,75	0,00	-	6,17	4,65	5,77	3,03	1,45	0,00
DET	4,11	9,05	6,43	7,51	5,88	-	7,41	2,33	7,69	12,12	14,49	7,69
INTJ	0,00	0,44	0,24	0,13	0,00	-	0,00	0,00	0,00	0,00	0,00	0,00
MD	0,00	0,00	0,00	0,00	0,00	-	0,00	0,00	0,00	0,00	0,00	0,00
PREP	8,56	4,86	8,81	4,83	0,00	-	8,64	9,30	7,69	3,03	7,25	7,69
PRON	2,74	3,97	0,71	2,28	0,00	-	4,94	0,00	2,88	0,00	4,35	0,00
NOM	35,62	31,13	30,00	29,89	23,53	-	20,99	30,23	29,81	39,39	33,33	38,46
VERB	16,78	16,78	14,05	14,21	35,29	-	20,99	16,28	18,27	9,09	15,94	15,38

Author Profiling in Social Media

PAN2013

Task Proposal

- In classical authorship attribution, we are given a closed set of candidate authors and are asked to identify which one of them is the author of an anonymous text. In addition, authorship attribution can distinguish between classes of authors, rather than individual authors, studying how language is shared by people. This helps in identifying profiling aspects such as: gender, age, native language, personality type, etc. Authorship profiling is a problem of growing importance in applications in forensics, security and marketing. For instance, from a forensic linguistics perspective being able to know what is the linguistic profile of a suspected text message (language used by a certain type of people) and identify characteristics (language as evidence) just by analyzing the text would certainly help considering suspects. Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products. We plan to deal with the problem of authorship profiling in social media being mainly interested in everyday language and how this reflects basic social and personality processes. Our starting point is the research carried out in the seminal work [1] where it was showed that statistical analysis of word usage in documents could be used to determine an author's gender, age, native language and personality type.
- In PAN-2013 we aim at considering the gender and age aspects of the authorship profiling problem, both in English and Spanish. So far research work in computational linguistics [1] and social psychology [2] has been carried out mainly for English and we believe interesting to investigate what are also the distinguishing features that may help in the gender and age classification task in a language different than English. In order to investigate the feasibility of organizing such task, we are in the process of assembling a corpus that will ultimately include tens of thousands of labeled examples. As age classes, we followed what was done in [1] and considered the three classes: 10s (13-17), 20s (23-27), and 30s (33-47). As for evaluation, we plan to use the standard precision, recall and F-score combined for both the age and gender classification aspects of the authorship profiling problem. Last, as case use we are discussing the possibility of including documents composed of chat lines of the authorship profiling task on Sexual Predator Identification that was organized at PAN-2012 [3], being the ground truth (in terms of gender and age) of victims and sexual predators already available.

[1] S. Argamon, M. Koppel, J. Pennebaker and J. Schler (2009), Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2): 119--123
[2] J. Pennebaker (2011). The secret life of pronouns: What our words say about us. New York: Bloomsbury Publishing, 2011
[3] G. Inches and F. Crestani (2012). Overview of the International Sexual Predator Identification Competition at PAN-2012. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, CLEF 2012 Evaluation Labs and Workshop -- Working Notes Papers, September 2012.

Co-organizers

Francisco Rangel

Autoritas Consulting

Paolo Rosso

Universitat Politècnica de València

Moshe Koppel

Bar-Ilan University

Efstathios Stamatatos

University of the Aegean

Giacomo Inches

University of Lugano

State of the Art

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Holmes & Meyerhoff, 2003 Burger & Henderson, 2006	Formal texts	-	Age and gender	
Koppel et al., 2003	Blogs	Simple lexical and syntactic functions	Gender: 80% accuracy	Self-labeling
Schler et al., 2006	Blogs	Stylistic features + content words with the highest information gain	Gender: 80% accuracy Age: 75% accuracy	
Goswami et al., 2009	-	Slang + sentence length	Gender: 89.18 accuracy Age: 80.32 accuracy	
Zhang & Zhang, 2010	Segments of blog		Gender: 72,10 accuracy	
Nguyen et al., 2011 y 2013	Blogs y Twitter		Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years	Manual labeling Age as continuous variable -> Logistic regresion
Peersman et al., 2011	Netlog			Self-labeling

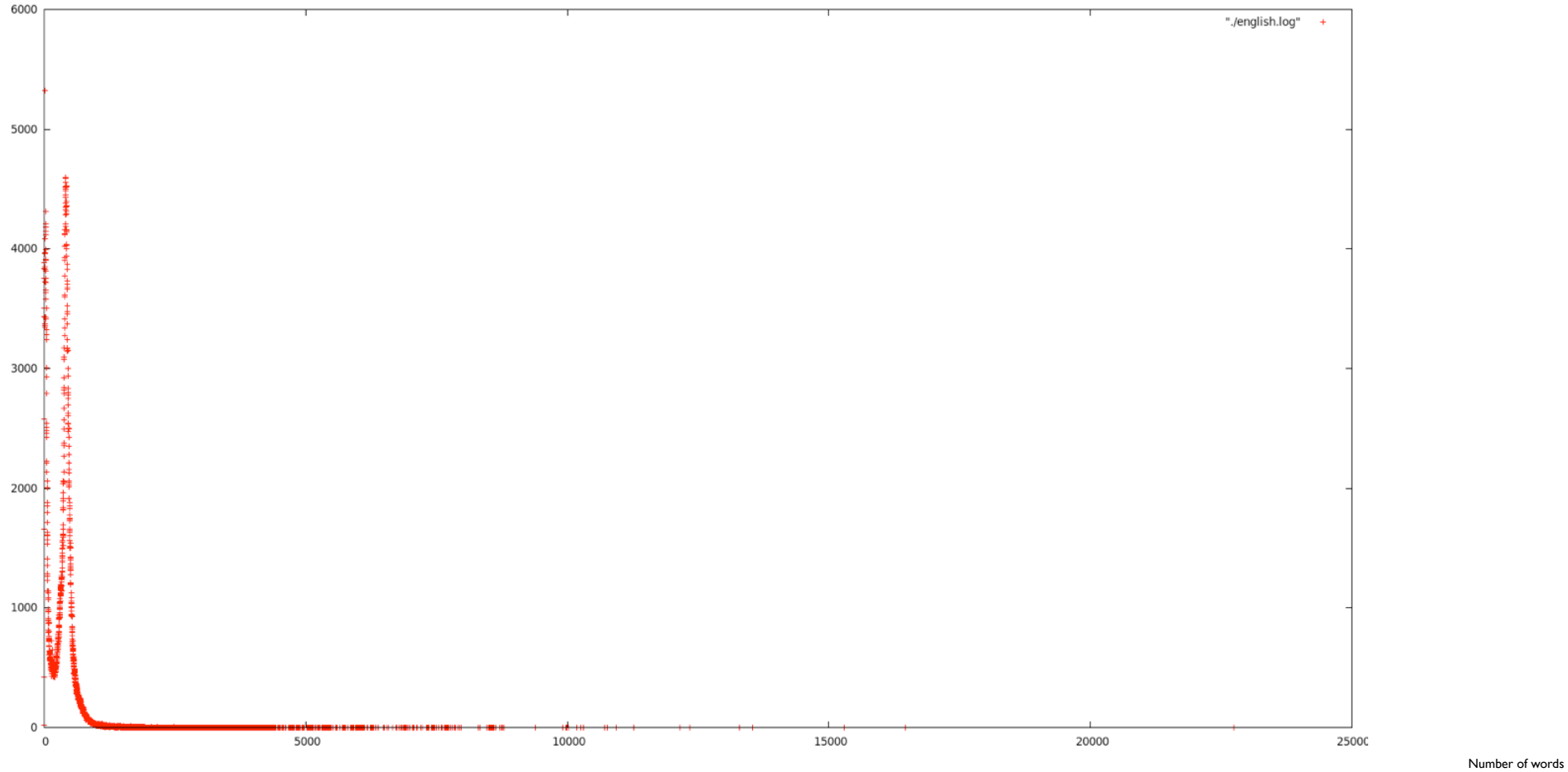
Information Retrieved

- Number of users crawled: 30,043,934
- Number of documents: 2,910,874 -> Not all users have a blog
- Number of words: 837,010,833

RAW DataSet Statistics

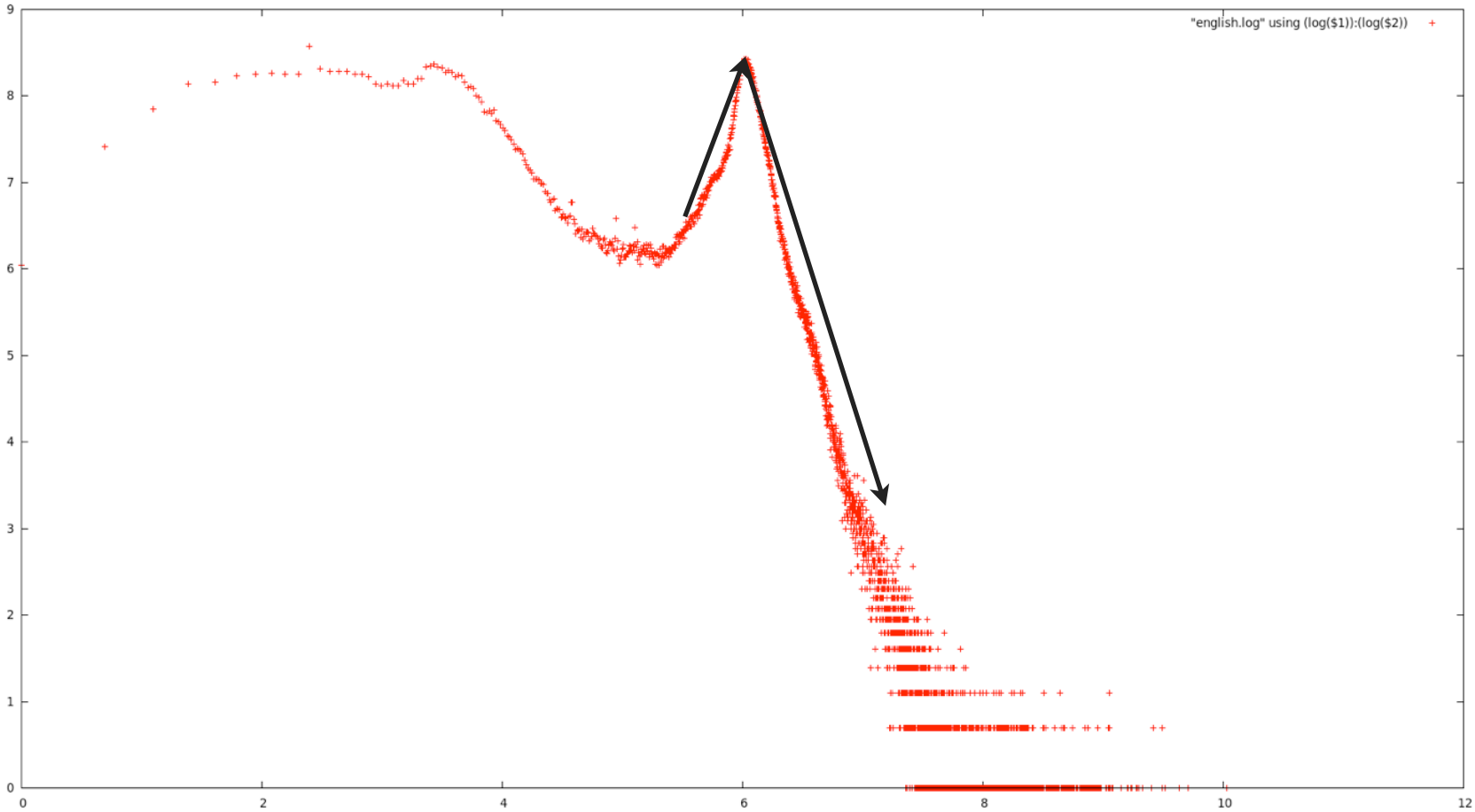
LANG	AGE	GENDER	NUM. DOCS / NUM. WORDS (AVERAGE / STANDAR DEVIATION)	NUM. DOCS / NUM. WORDS (AVERAGE / STANDAR DEVIATION)	NUM. DOCS / NUM. WORDS (AVERAGE / STANDAR DEVIATION)
ENGLISH	10s	MALE	32,333 / 13,246,053 (409 / 123)	67,394 / 26,644,067 (395 / 149)	986,657 / 331,005,073 (335 / 208)
		FEMALE	35,061 / 13,398,014 (382 / 174)		
	20s	MALE	192,598 / 49,012,979 (254 / 291)	355,921 / 93,973,707 (264 / 263)	
		FEMALE	163,323 / 44,960,728 (275 / 230)		
	30s	MALE	308,261 / 111,604,811 (362 / 212)	563,342 / 210,387,299 (373 / 212)	
		FEMALE	255,081 / 98,782,488 (387 / 137)		
SPANISH	10s	MALE	2,839 / 422,274 (148 / 988)	11,138 / 1,470,356 (132 / 673)	231,006 / 40,737,249 (176 / 832)
		FEMALE	8,299 / 1,048,082 (126 / 544)		
	20s	MALE	60,231 / 10,285,742 (170 / 795)	121,610 / 20,385,372 (167 / 646)	
		FEMALE	61,379 / 10,099,630 (164 / 494)		
	30s	MALE	50,932 / 11,139,064 (1218 / 1406)	98,258 / 18,881,521 (192 / 1043)	
		FEMALE	47,326 / 7,742,457 (163 / 512)		

English Distribution



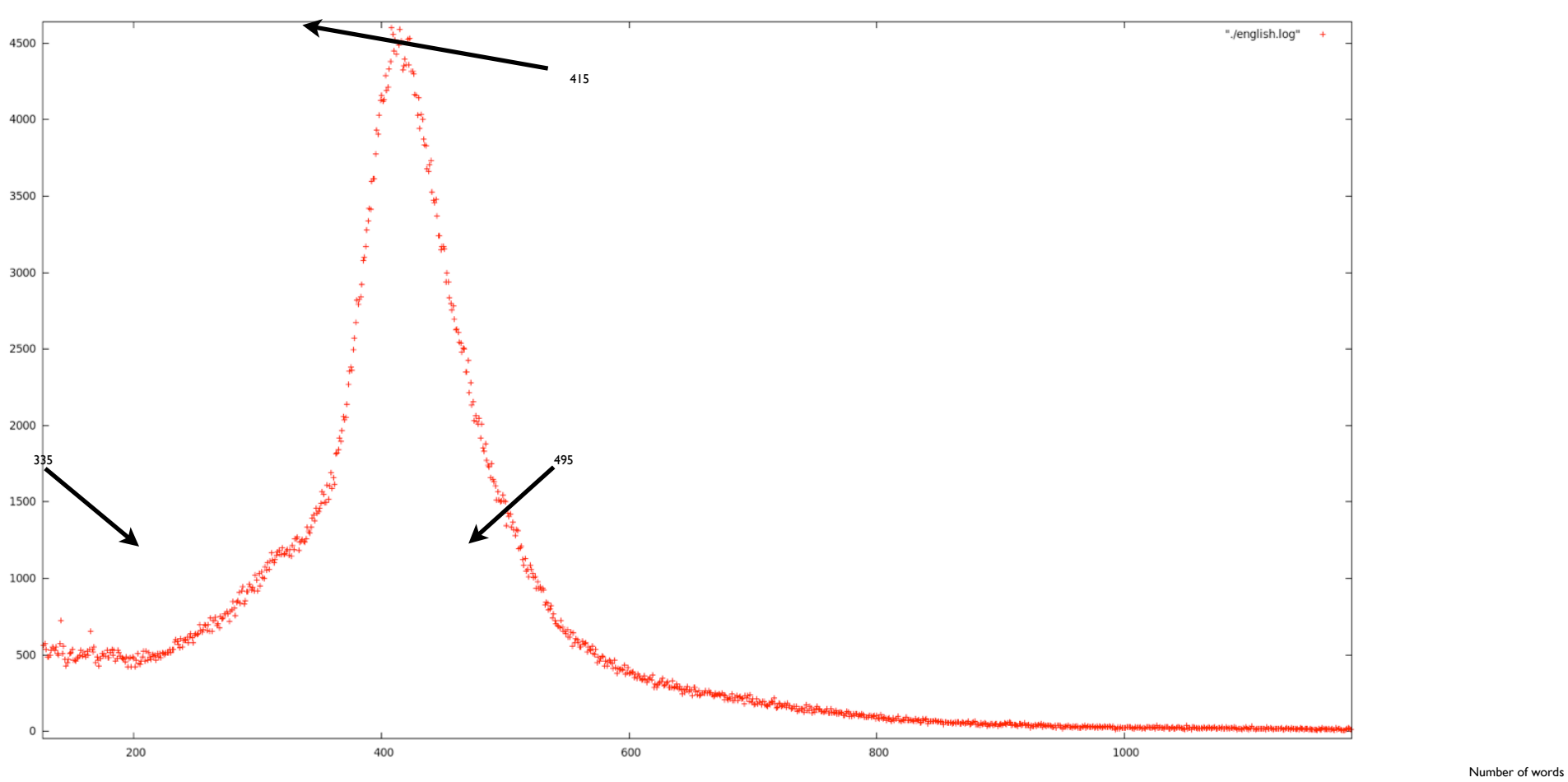
MIN	MAX	AVG	STD
0	22,736	335	208

English Distribution log-log



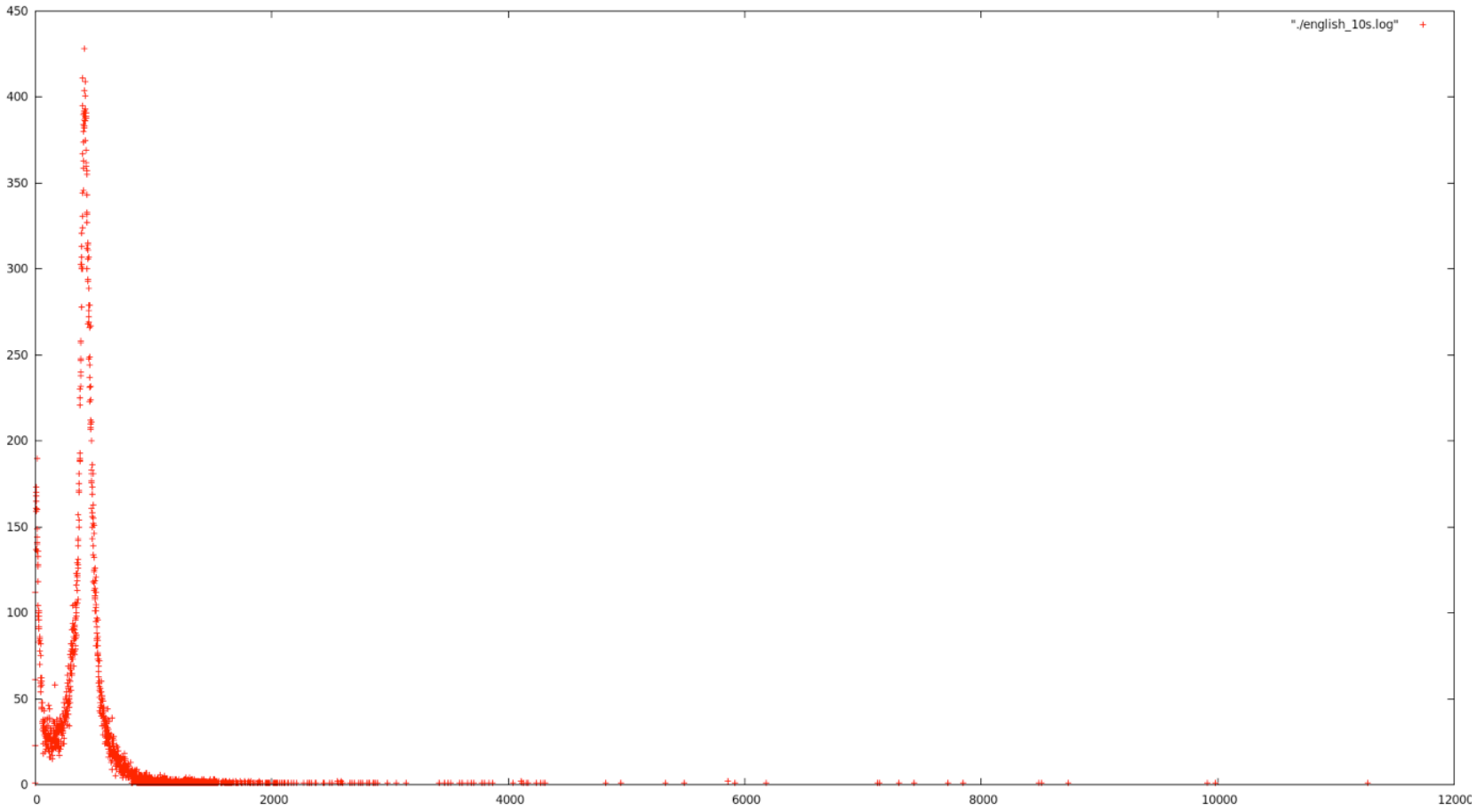
- The log-log representation shows how the previous distribution has a long tail component, specifically in two cases, before the point of maximun frequency and one more after this.
- We can use this property to select minimum and maximum number of words that the posts must have.

English Distribution zoomed



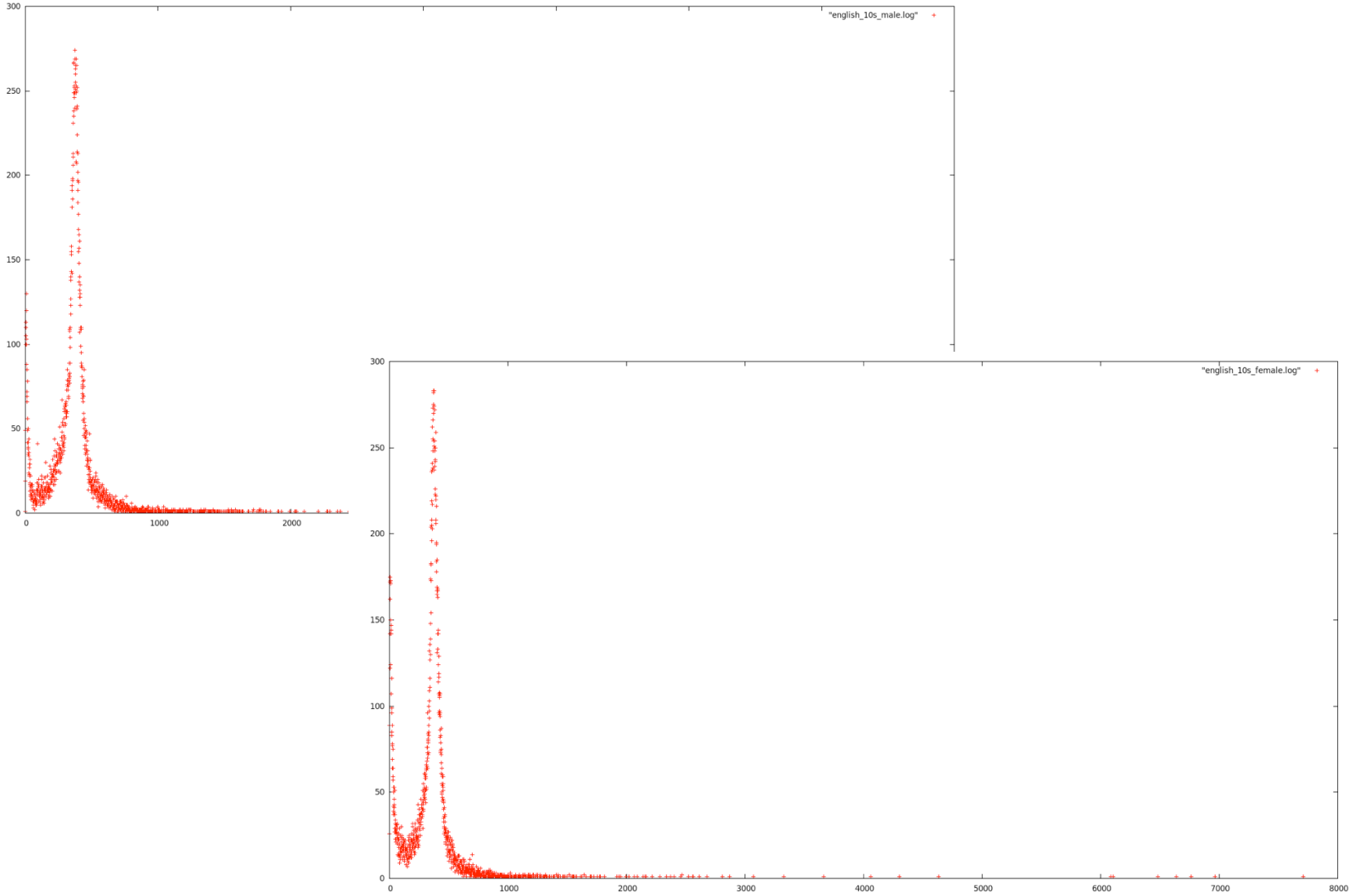
- If we zoom the distribution, we can observe a gassian like distribution, with its maximum on the value 415. If we web the global average as the point to obtain the standard deviation, we have the values printed in the chart

English Distribution 10s

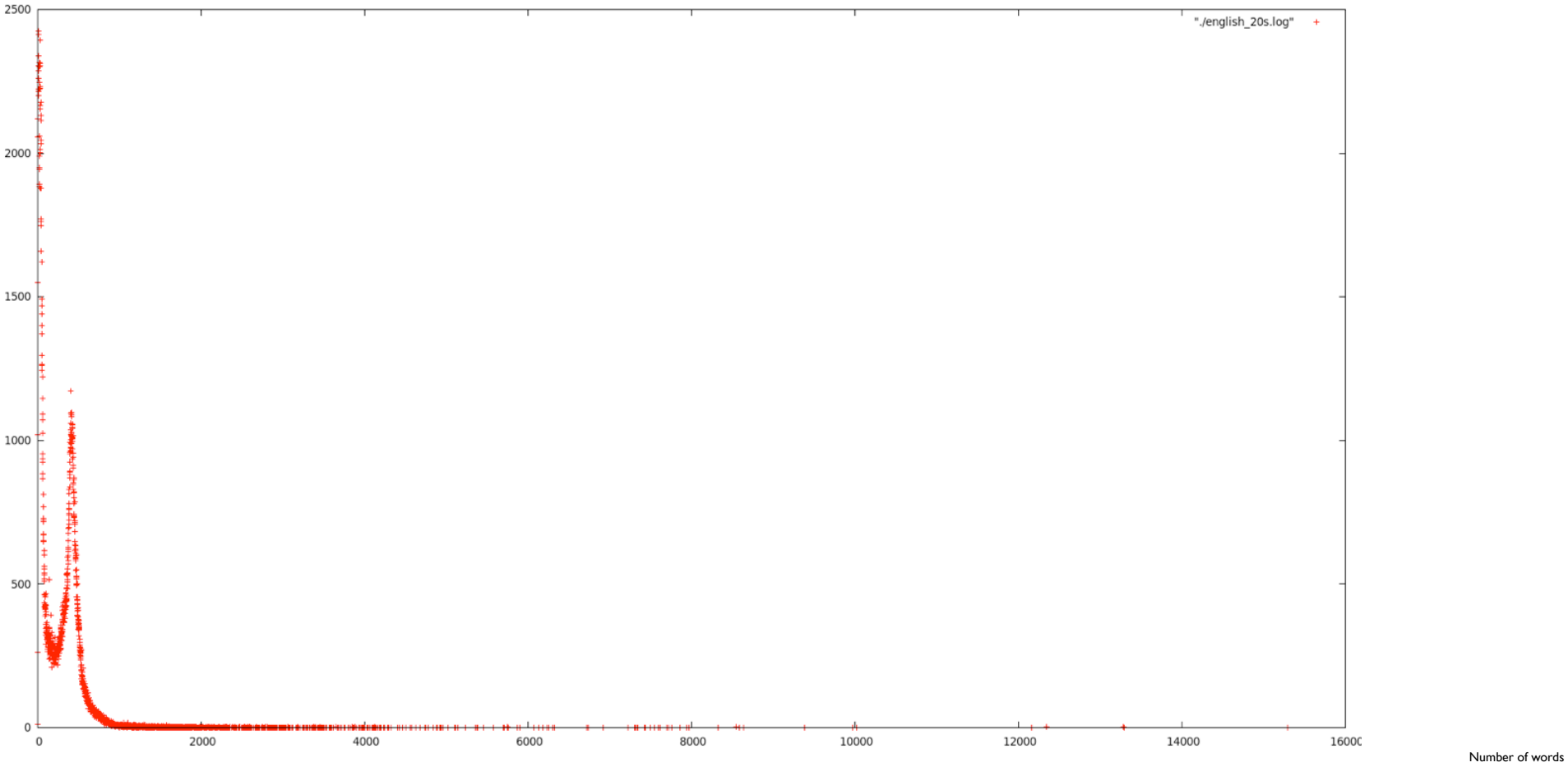


MIN	MAX	AVG	STD
0	11,267	395	149

English Distribution 10s Male-Female

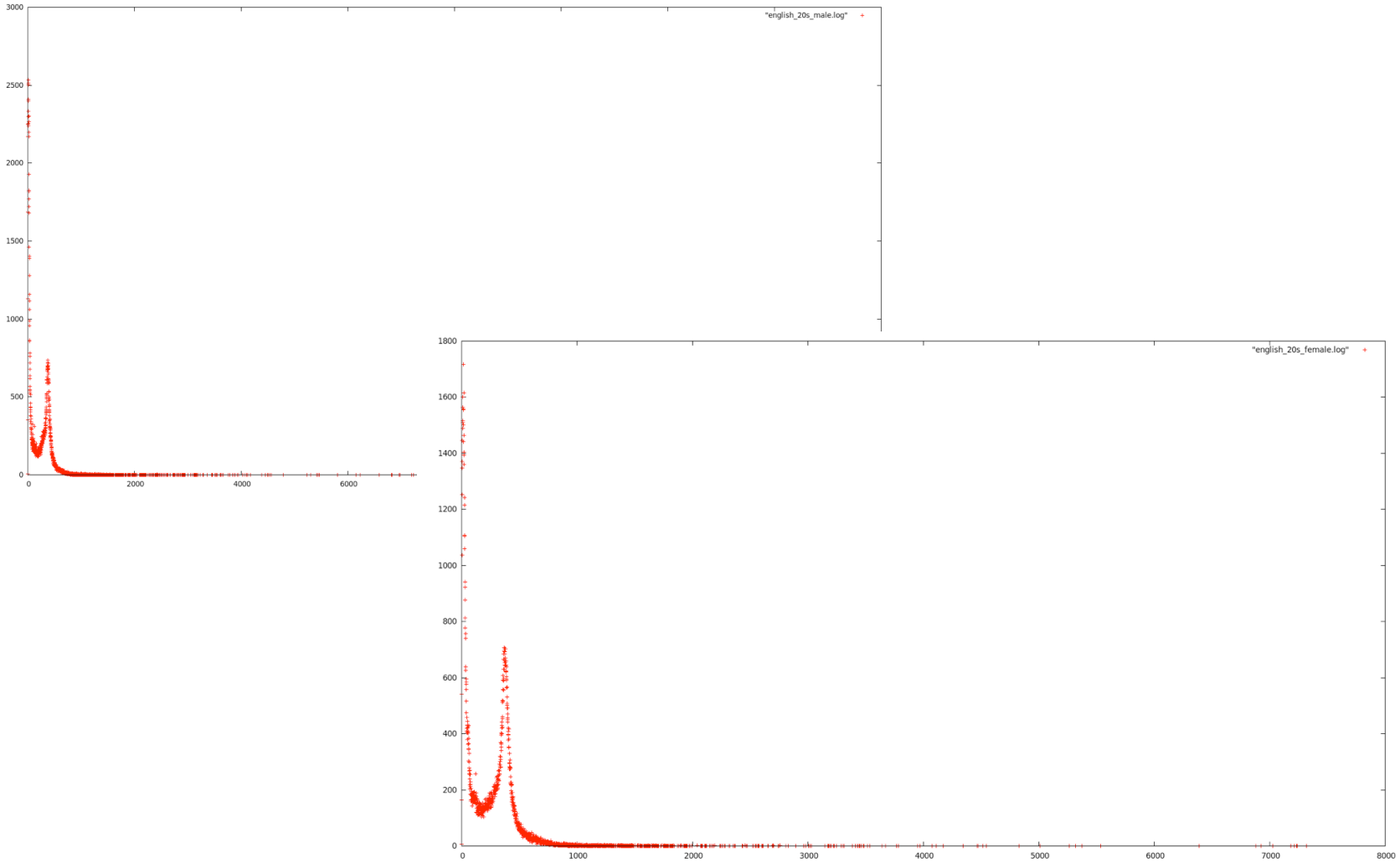


English Distribution 20s

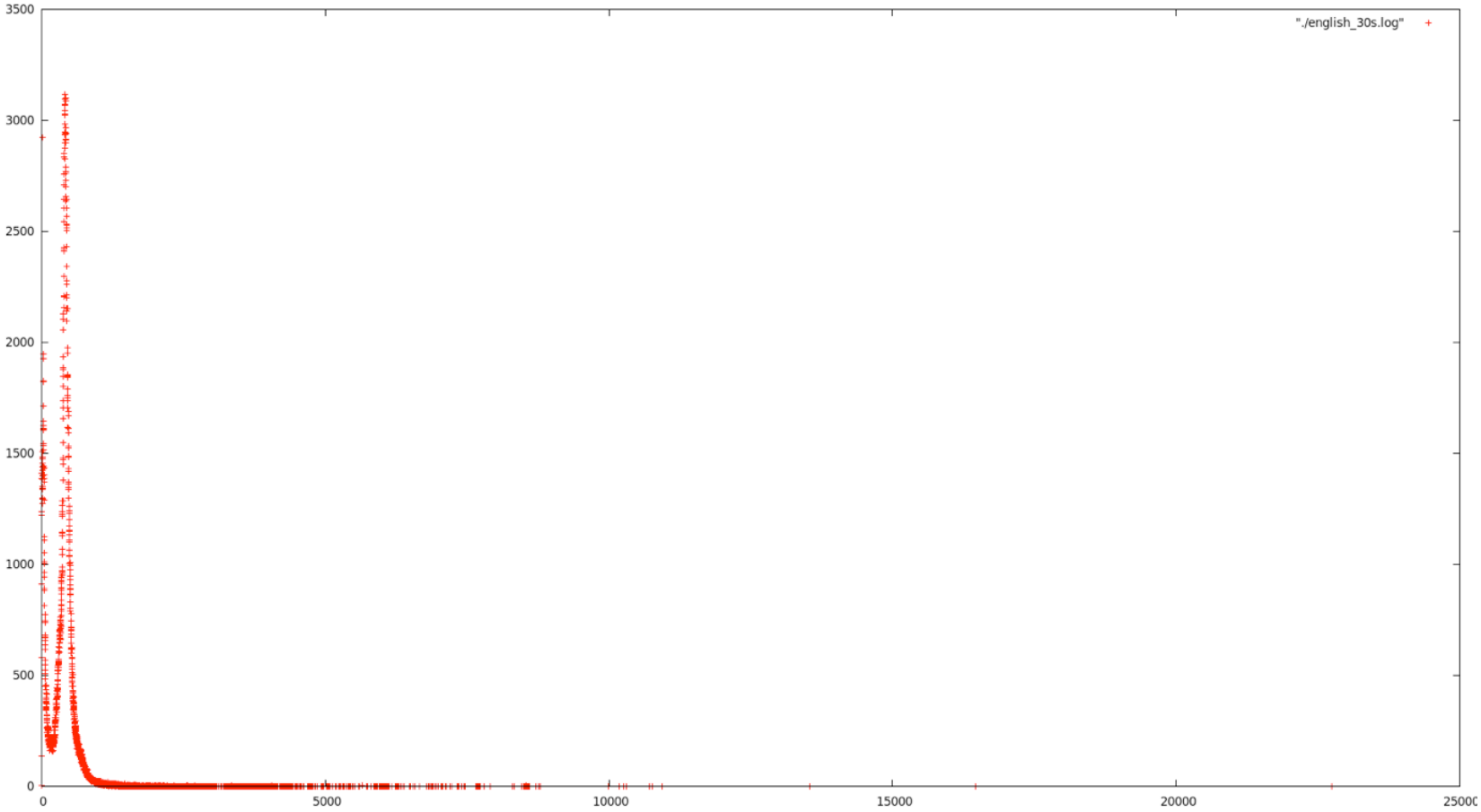


MIN	MAX	AVG	STD
0	15,292	264	263

English Distribution 20s Male-Female

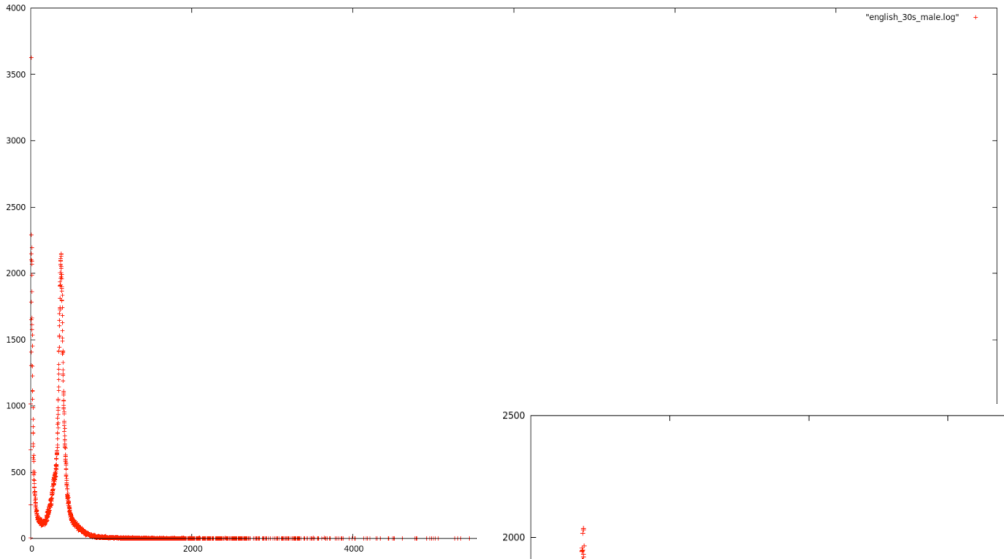


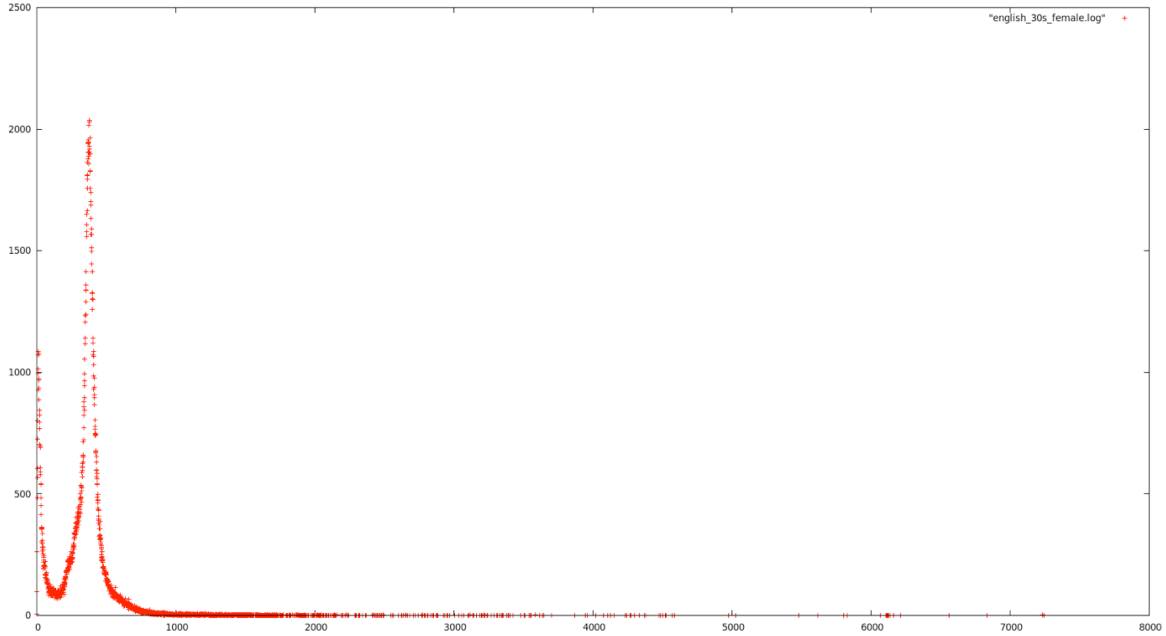
English Distribution 30s



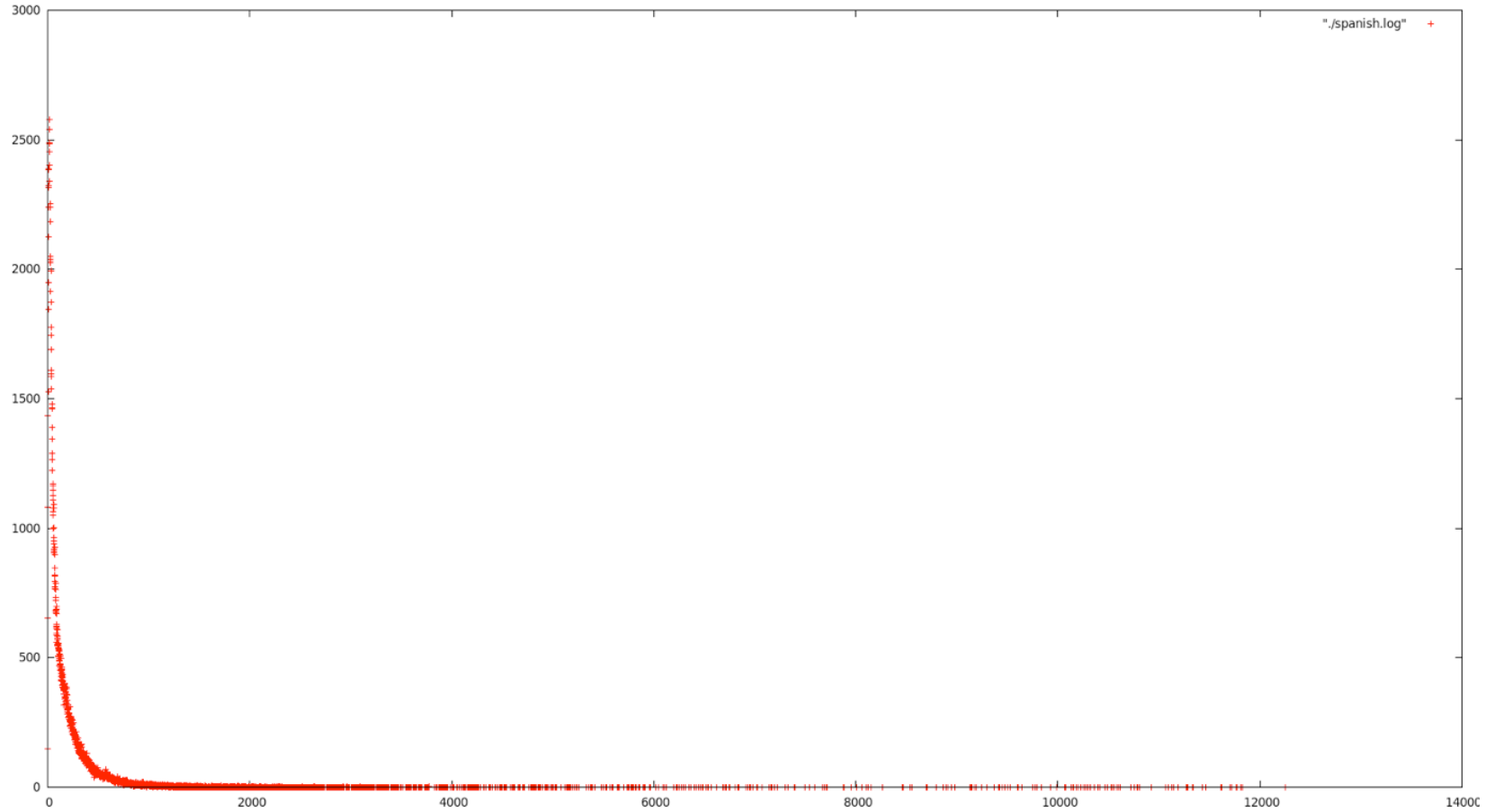
MIN	MAX	AVG	STD
0	22,735	373	177

English Distribution 30s Male-Female



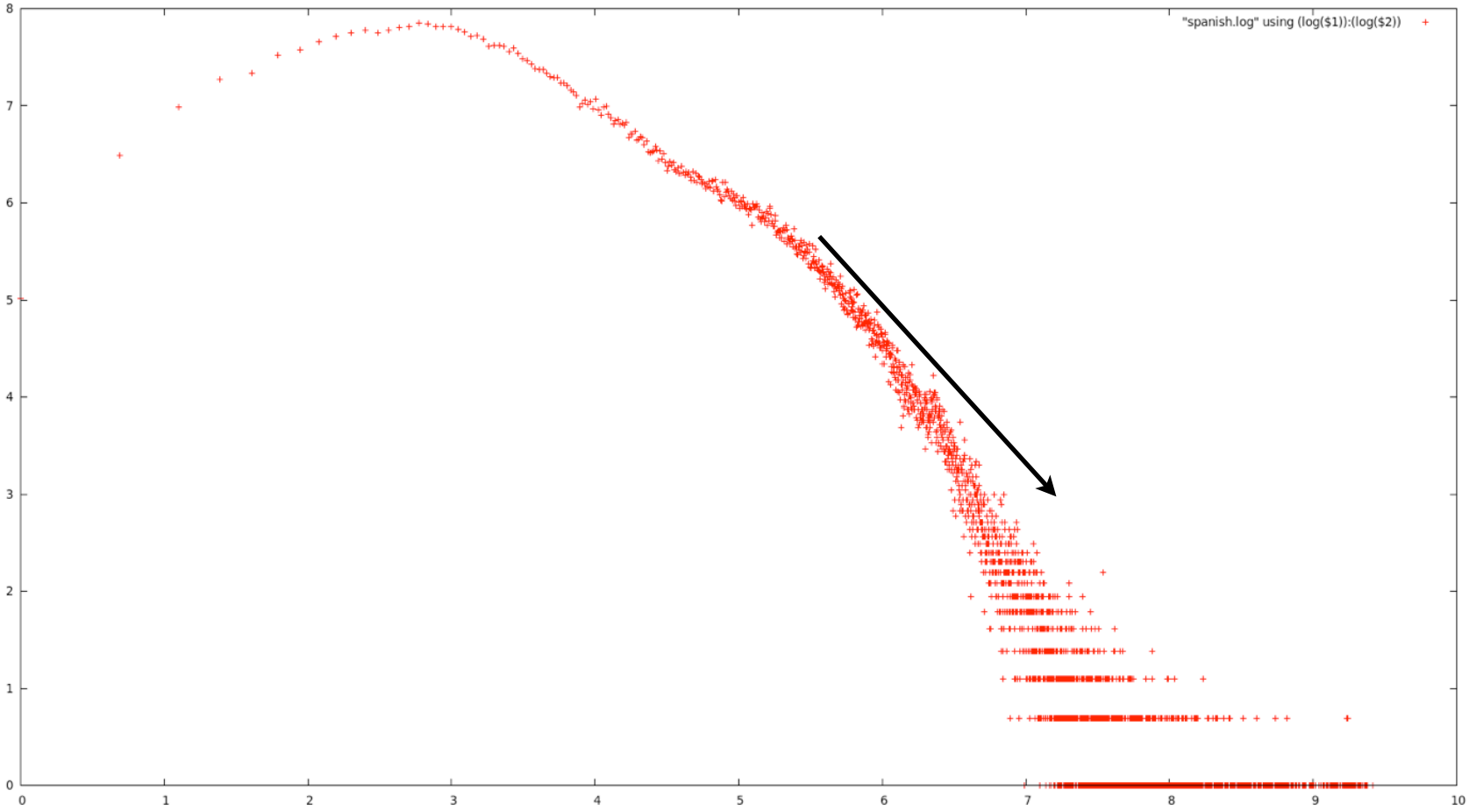


Spanish Distribution

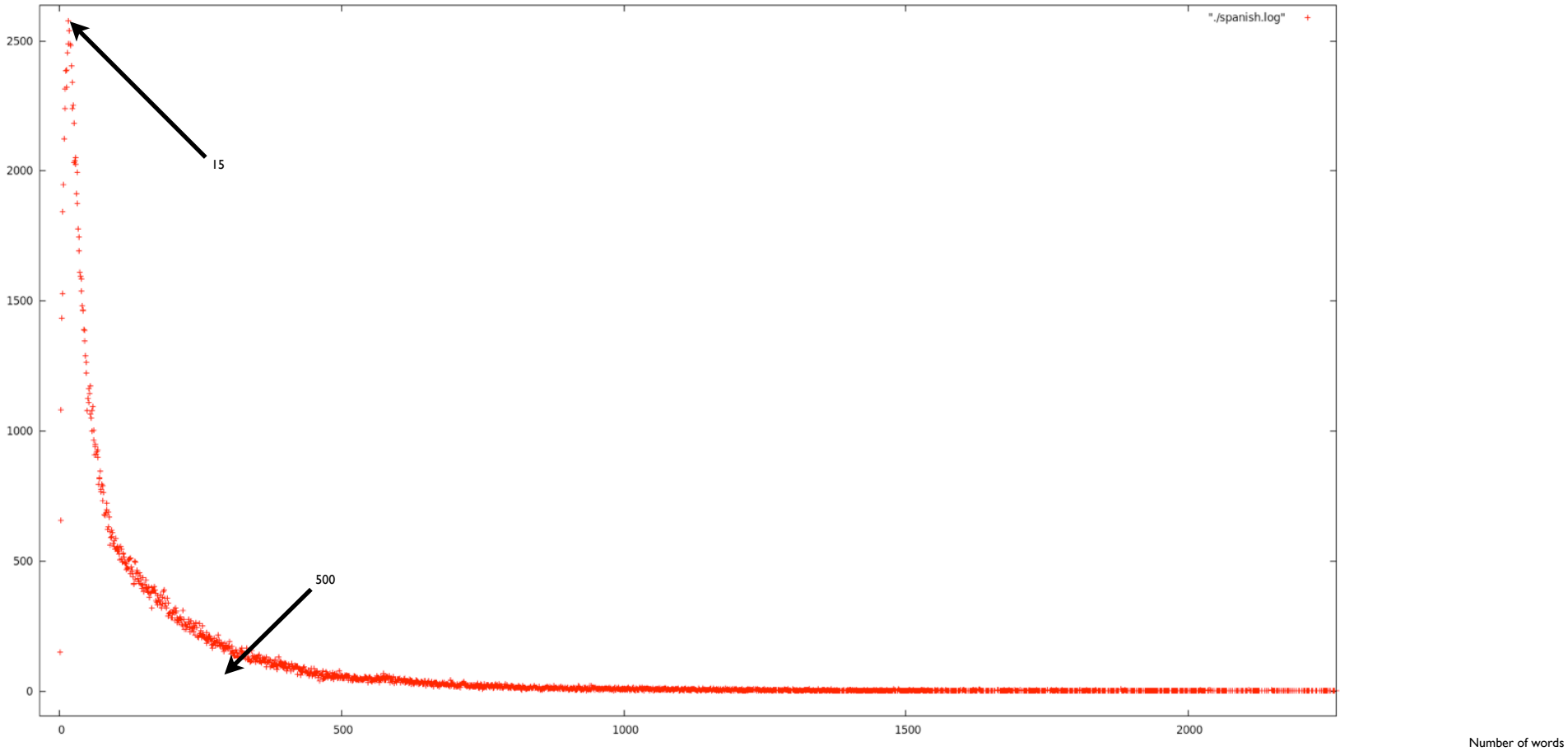


MIN	MAX	AVG	STD
0	12,246	176	832

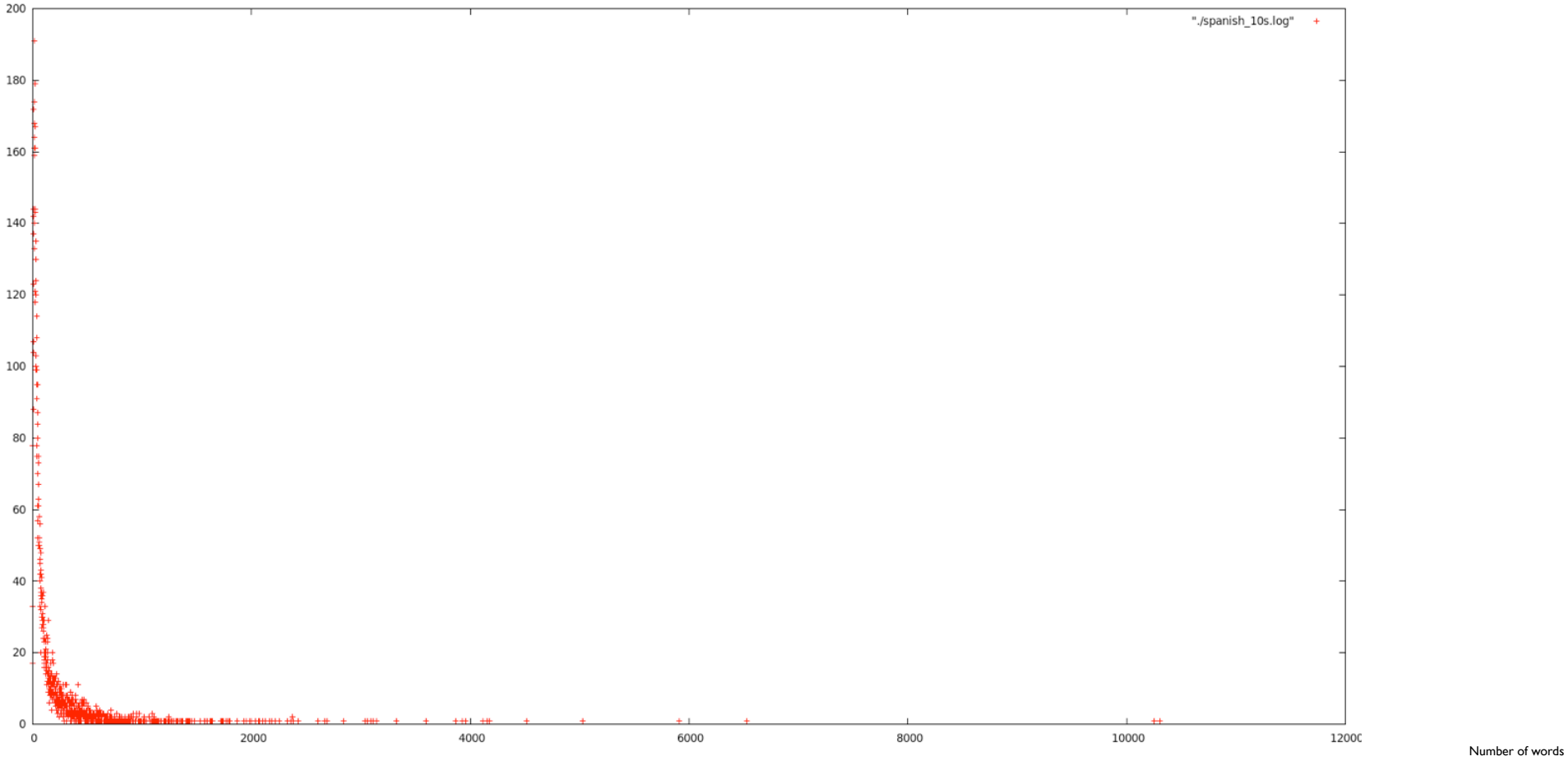
Spanish Distribution log-log



Spanish Distribution zoomed

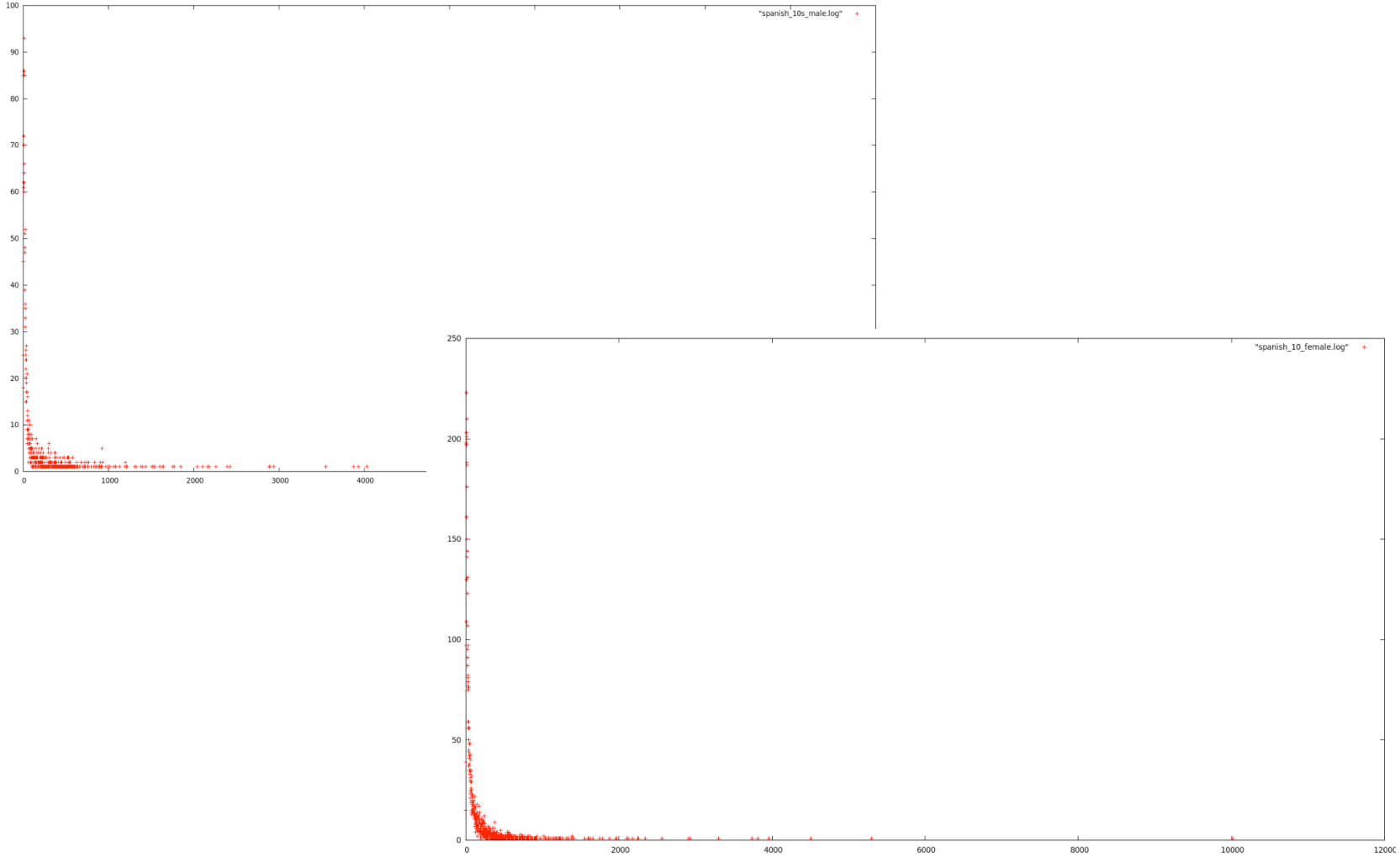


Spanish Distribution 10s

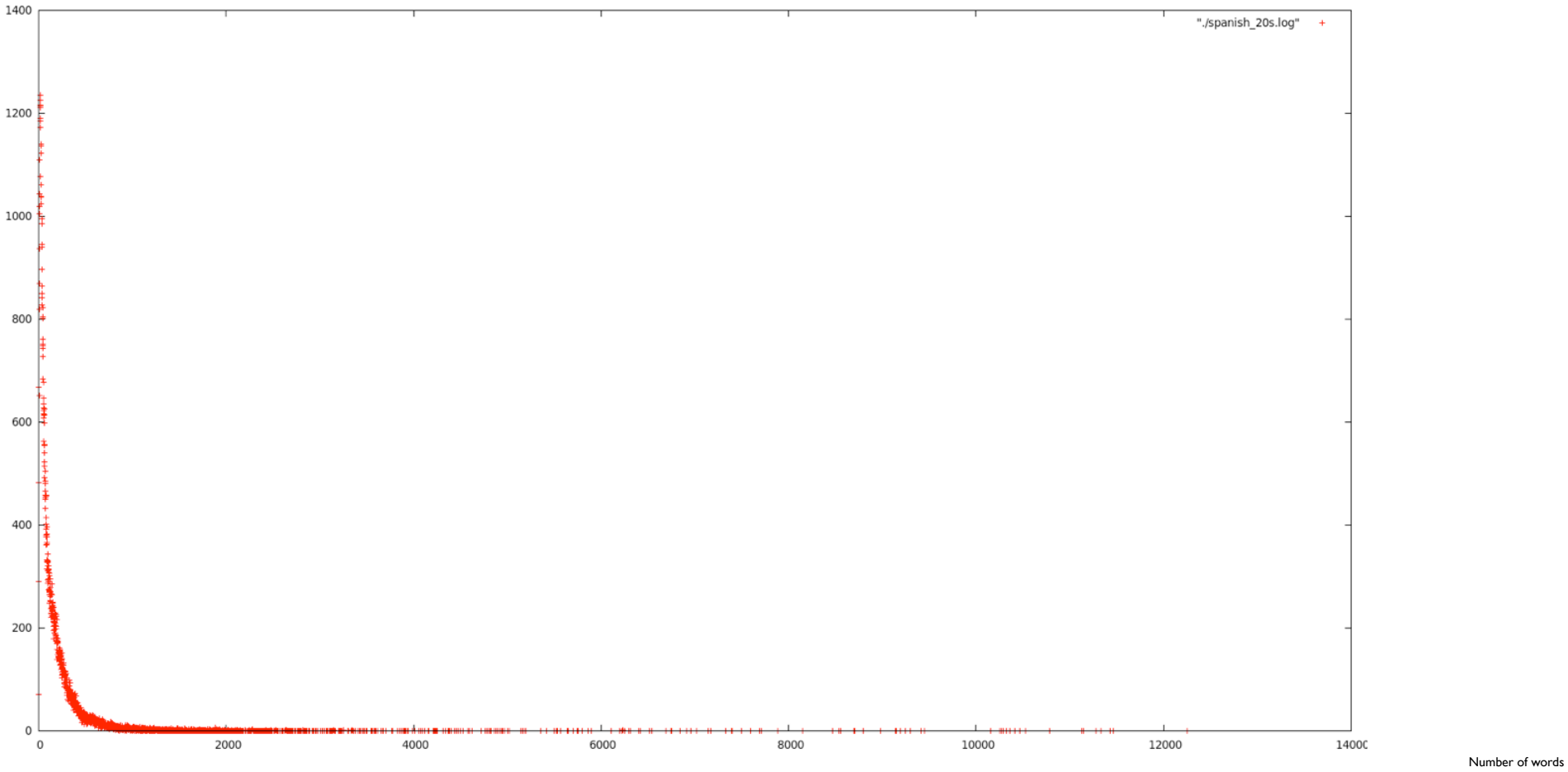


MIN	MAX	AVG	STD
1	10,301	132	673

Spanish Distribution 10s Male-Female

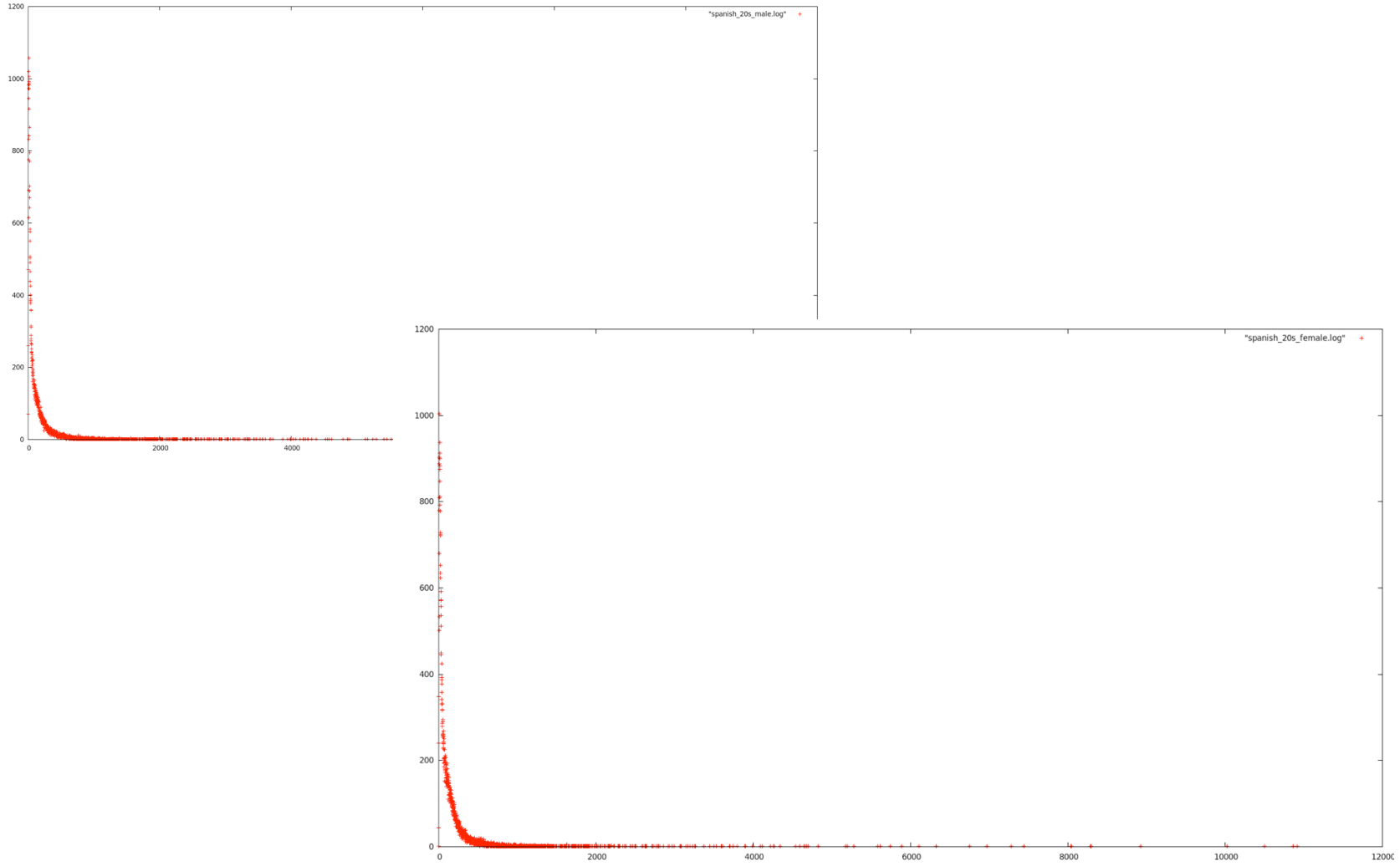


Spanish Distribution 20s

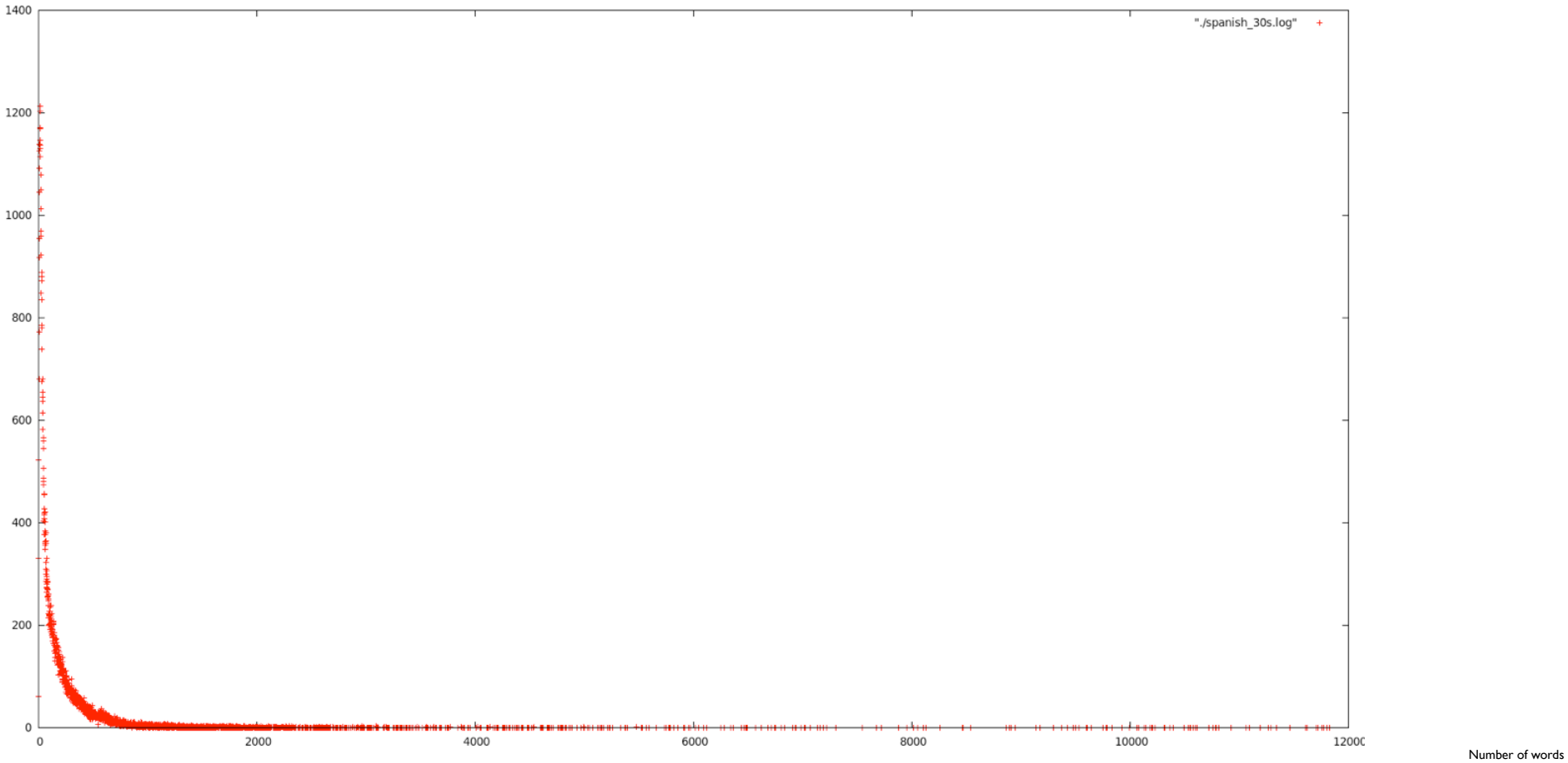


MIN	MAX	AVG	STD
0	12,246	167	646

Spanish Distribution 20s Male-Female

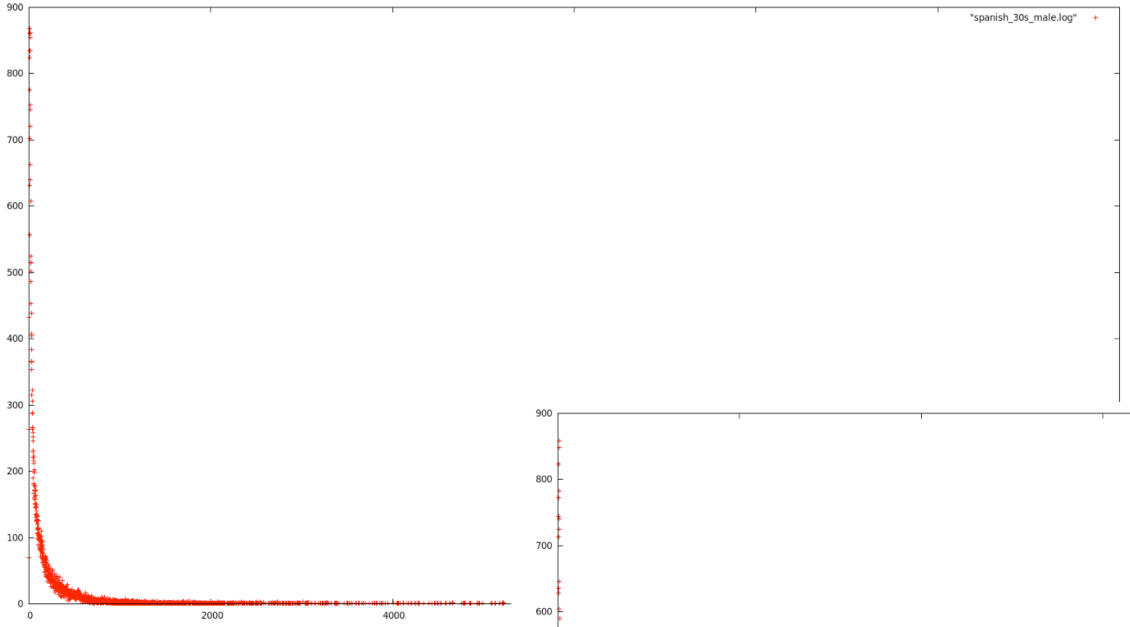


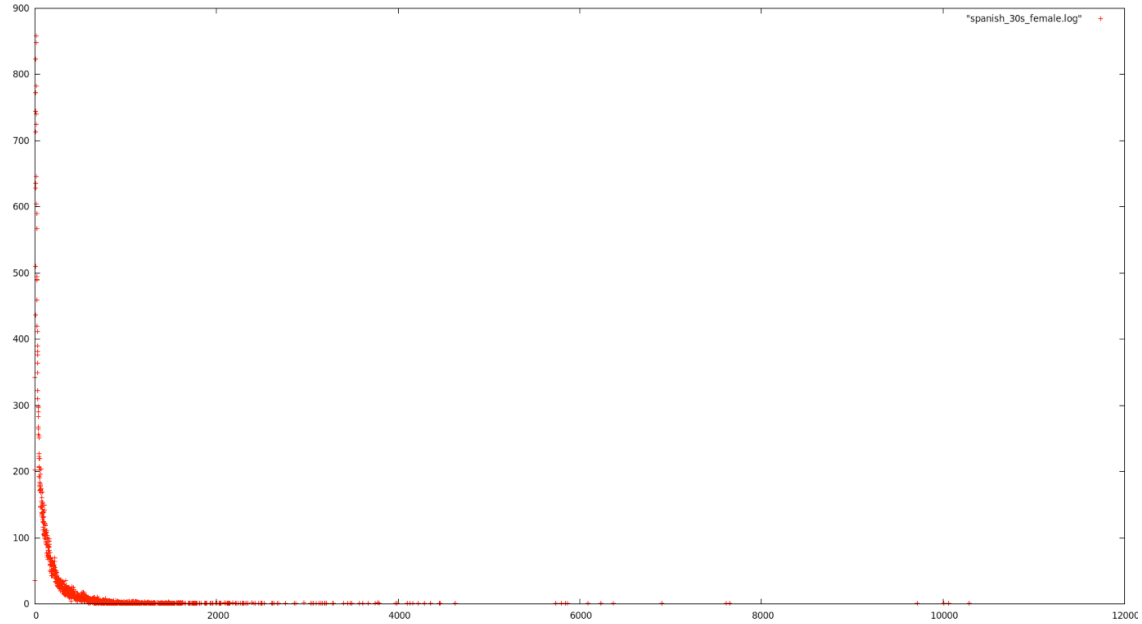
Spanish Distribution 30s



MIN	MAX	AVG	STD
1	11,824	192	1043

Spanish Distribution 30s Male-Female





Posts grouped by author

LANG	AGE	GENDER	NUM. AUTHORS / NUM. WORDS (AVERAGE / STANDARD DEVIATION)	NUM. AUTHORS / NUM. WORDS (AVERAGE / STANDARD DEVIATION)	NUM. AUTHORS / NUM. WORDS (AVERAGE / STANDARD DEVIATION)
EN	10s	MALE	12,642 / 13,430,304 (1,062 / 3,404)	27,086 / 27,013,621 (997 / 3,363)	400,735 / 336,668,085 (840 / 3,400)
		FEMALE	14,444 / 13,583,317 (940 / 3,315)		
	20s	MALE	83,668 / 50,080,342 (598 / 3,916)	150,204 / 95,952,931 (638 / 3,842)	
		FEMALE	66,536 / 45,872,589 (689 / 3,756)		
	30s	MALE	122,633 / 113,385,388 (924 / 3,280)	223,445 / 213,701,533 (956 / 3,160)	
		FEMALE	100,812 / 100,316,145 (995 / 3,020)		
ES	10s	MALE	2,111 / 431,042 (204 / 3,664)	8,045 / 1,493,898 (185 / 2,209)	127,760 / 40,546,718 (317 / 5,828)
		FEMALE	5,934 / 1,062,856 (179 / 1,618)		
	20s	MALE	35,306 / 10,322,408 (292 / 2,980)	69,090 / 20,374,555 (294 / 3,547)	
		FEMALE	33,784 / 10,052,147 (297 / 2,980)		
	30s	MALE	27,311 / 10,969,560 (401 / 11,600)	50,625 / 18,678,265 (368 / 8,590)	
		FEMALE	23,314 / 7,708,705 (330 / 4,298)		

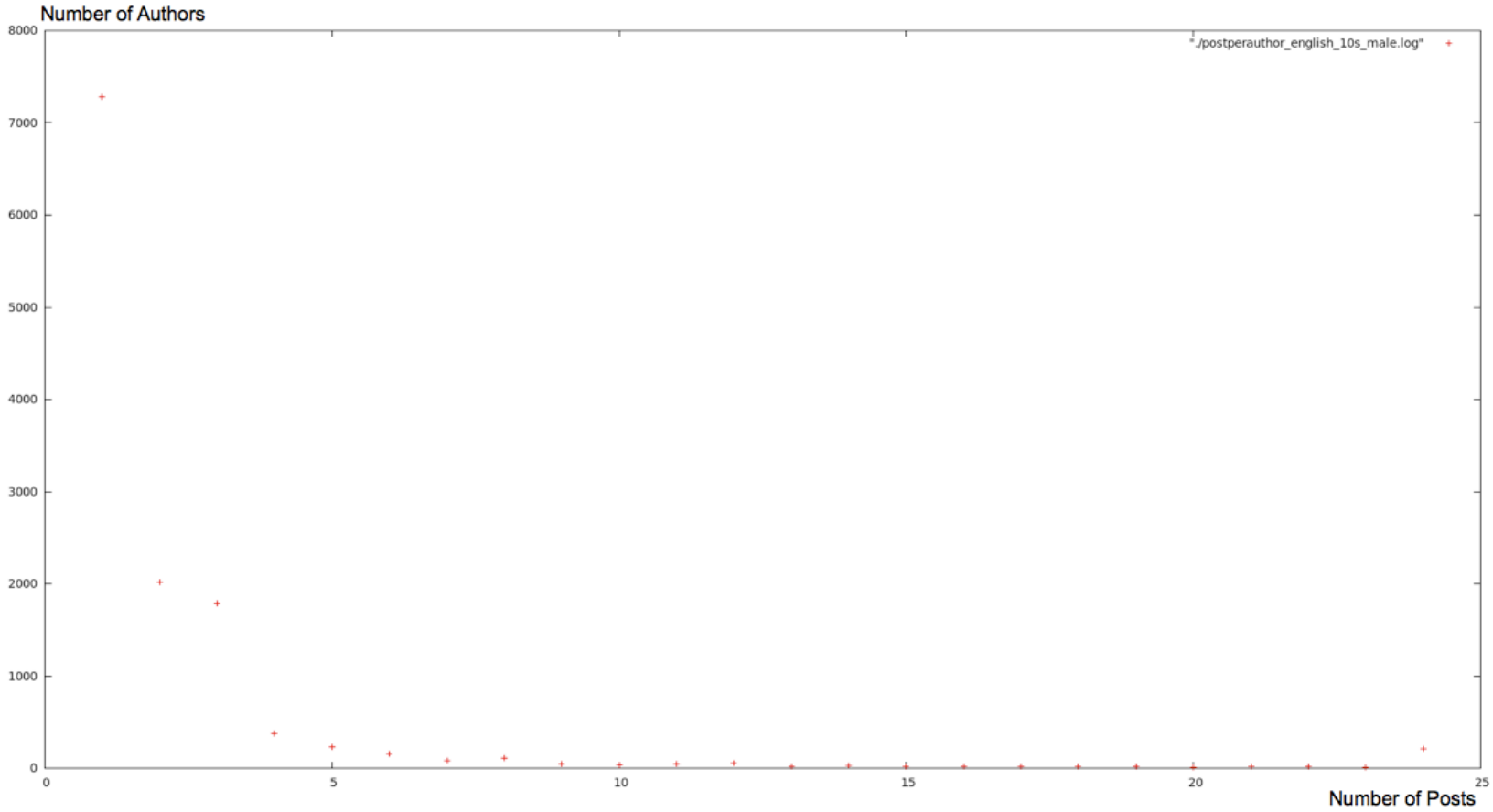
Posts per author: Distribution

N. POSTS	ENGLISH						SPANISH					
	10s		20s		30s		10s		20s		30s	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
1	7,280	8,635	57,407	43,247	74,357	58,244	1,779	4,814	28,143	25,543	21,626	17,267
2	2,020	2,347	10,546	8,820	18,270	16,135	195	676	3,520	4,011	2,657	2,767
3	1,785	1,799	6,251	5,946	14,324	13,444	59	206	1,260	1,466	935	1,070
4	378	448	2,122	1,910	3,894	3,314	26	90	573	752	478	501
5	228	246	1,309	1,216	2,062	2,035	15	48	371	462	292	352
6	159	172	975	905	1,683	1,406	11	32	258	308	225	215
7	80	93	619	596	1,056	786	6	14	196	217	145	160
8	112	92	571	567	1,082	888	7	8	144	173	104	165
9	51	59	446	394	635	471	5	9	141	115	84	97
10	36	47	466	348	551	393	30	6	88	103	78	68
11	43	46	299	218	388	296	30	7	83	83	64	71
12	53	35	204	195	400	274	0	8	59	66	52	41
13	23	3	171	159	260	196	0	3	36	52	43	56
14	30	18	157	140	261	171	0	2	47	30	39	43
15	17	28	132	127	190	188	0	1	32	49	38	42
16	20	24	115	95	203	165	1	1	29	21	39	26
17	17	14	87	67	147	116	0	1	37	18	31	27
18	11	12	94	88	110	91	1	3	23	25	28	20
19	19	9	58	61	123	94	0	1	21	20	13	12
20	19	10	71	59	113	107	0	1	15	25	15	22
21	22	11	81	65	93	92	0	3	17	28	19	11
22	20	10	59	57	91	79	0	0	18	10	16	16
23	11	10	58	46	105	76	0	0	13	12	10	7
24	213	249	1,374	1,210	2,235	1,751	0	0	182	195	280	258

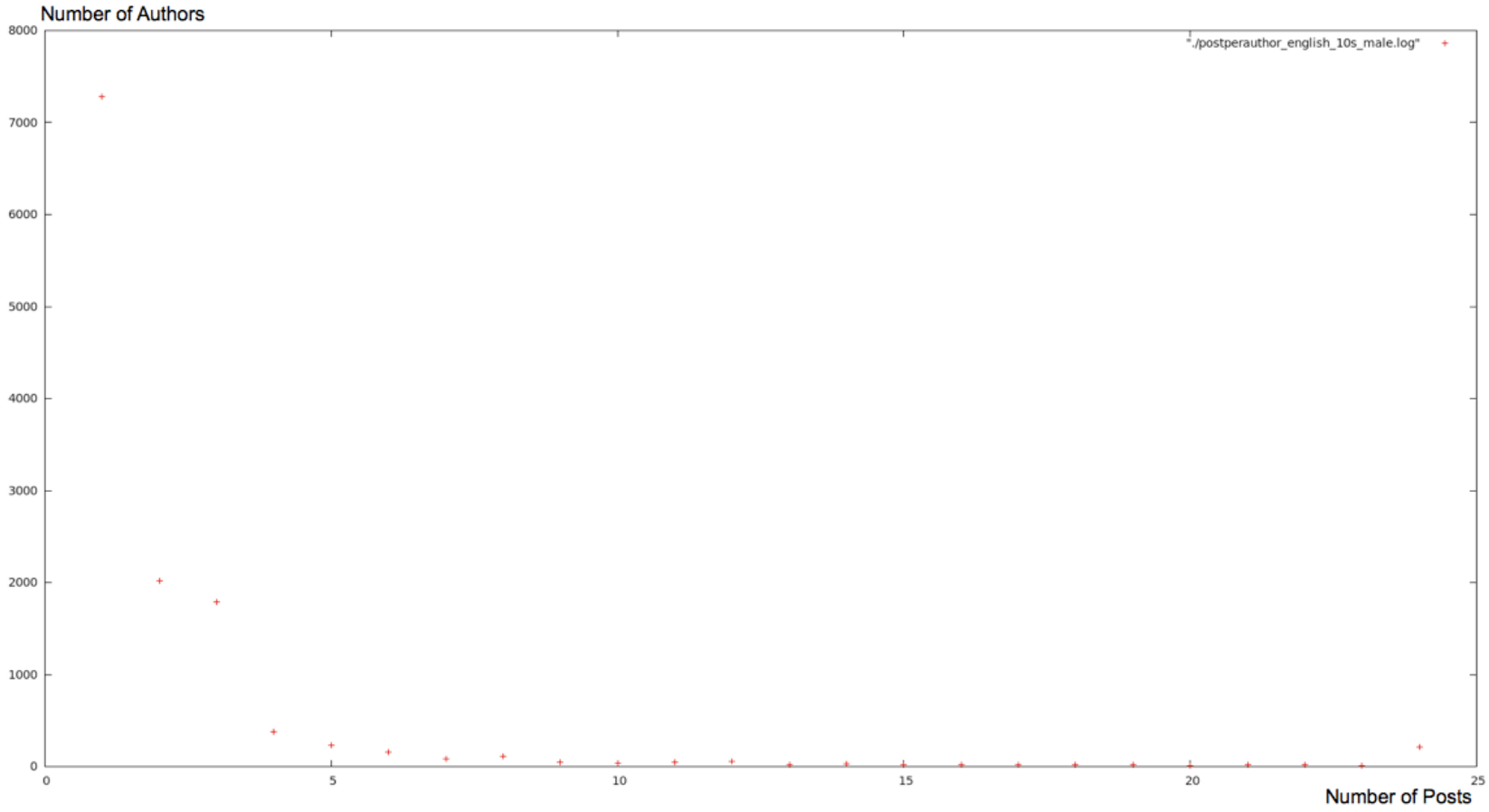
Posts per author: Distribution

N. POSTS	ENGLISH						SPANISH					
	10s		20s		30s		10s		20s		30s	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
N.AUT<5	11,463	13,229	76,326	59,923	110,855	91,137	2,059	5,786	33,496	31,722	25,696	21,605
N.AUT>=5	1,184	1,188	5,973	5,404	9,555	7,926	52	148	1,810	2,012	1,615	1,709
% TRAIN.	9.36%	8.24%	7.26%	8.27%	7.94%	8.00%	2.46%	2.49%	5.13%	5.96%	5.91%	7.33%
N.POST<5	19,327	21,748	112,285	92,445	179,765	154,277	2,525	7,384	43,110	43,281	33,117	29,775
N.POST>=	14,267	14,153	54,007	48,202	85,381	69,328	381	1,179	18,929	20,398	19,099	19,115
% TRAIN.	42,47%	39.42%	32.48%	34.27%	32.20%	31.00%	13.11%	13.77%	30.51%	32.03%	36.58%	39.10%
N.AUT<8	11,930	13,740	79,229	62,640	115,656	95,364	2,091	5,880	34,321	32,759	26,358	22,332
N.AUT>=8	717	677	3,070	2,687	4,754	3,699	20	54	985	1,025	953	982
% TRAIN.	5.67%	4.70%	3.73%	4.11%	3.95%	3.73%	0.95%	0.91%	2.79%	3.03%	3.49%	4.21%
N.POST<8	20,846	23,436	122,473	102,052	197,260	168,220	2,638	7,679	46,035	46,653	35,487	32,190
N.POST>=	11,613	11,240	37,279	32,520	57,581	45,215	198	649	14,154	14,721	15,274	14,945
% TRAIN.	35.78%	32.41%	23.34%	24.17%	22.59%	21.18%	6.98%	7.79%	23.52%	23.99%	30.09%	31.71%
N.AUT<10	12,093	13,891	80,246	63,601	117,373	96,723	2,103	5,897	34,606	33,047	26,546	22,594
N.AUT>=1	554	526	2,053	1,726	3,037	2,340	8	37	700	737	765	720
% TRAIN.	4.38%	3.65%	2.50%	2.64%	2.52%	2.36%	0.38%	0.62%	1.98%	2.18%	2.80%	3.09%
N.POST<1	22,196	24,698	131,050	110,129	211,626	179,558	2,734	7,819	48,451	49,067	37,070	34,378
N.POST>=	10,258	9,973	28,697	24,438	43,210	33,872	97	504	11,733	12,302	13,686	12,752
% TRAIN.	31.61%	28.76%	17.96%	18.16%	16.96%	15.87%	3.43%	6,06%	19.50%	20.05%	26.96%	27.06%

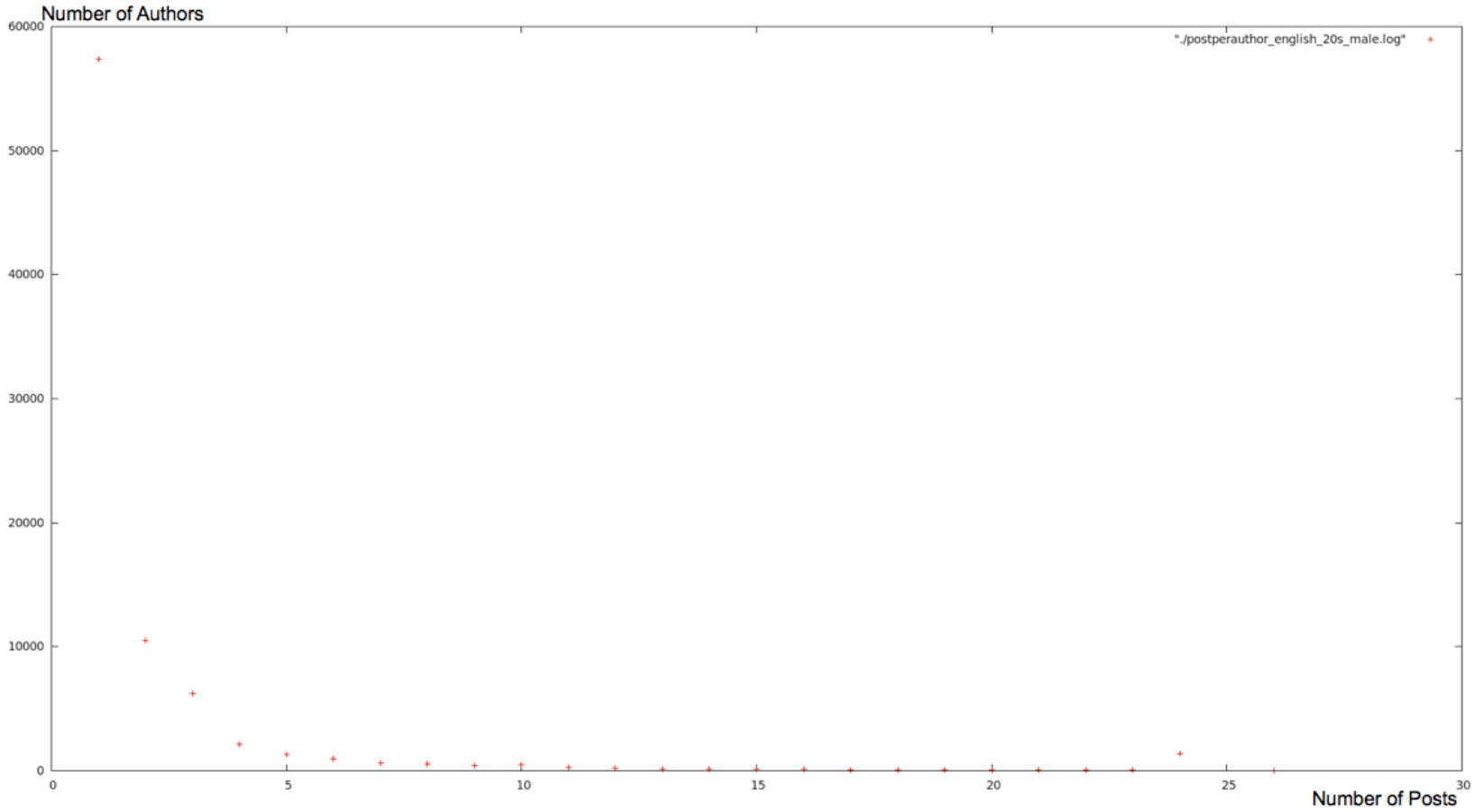
P.p.a.: English 10s male



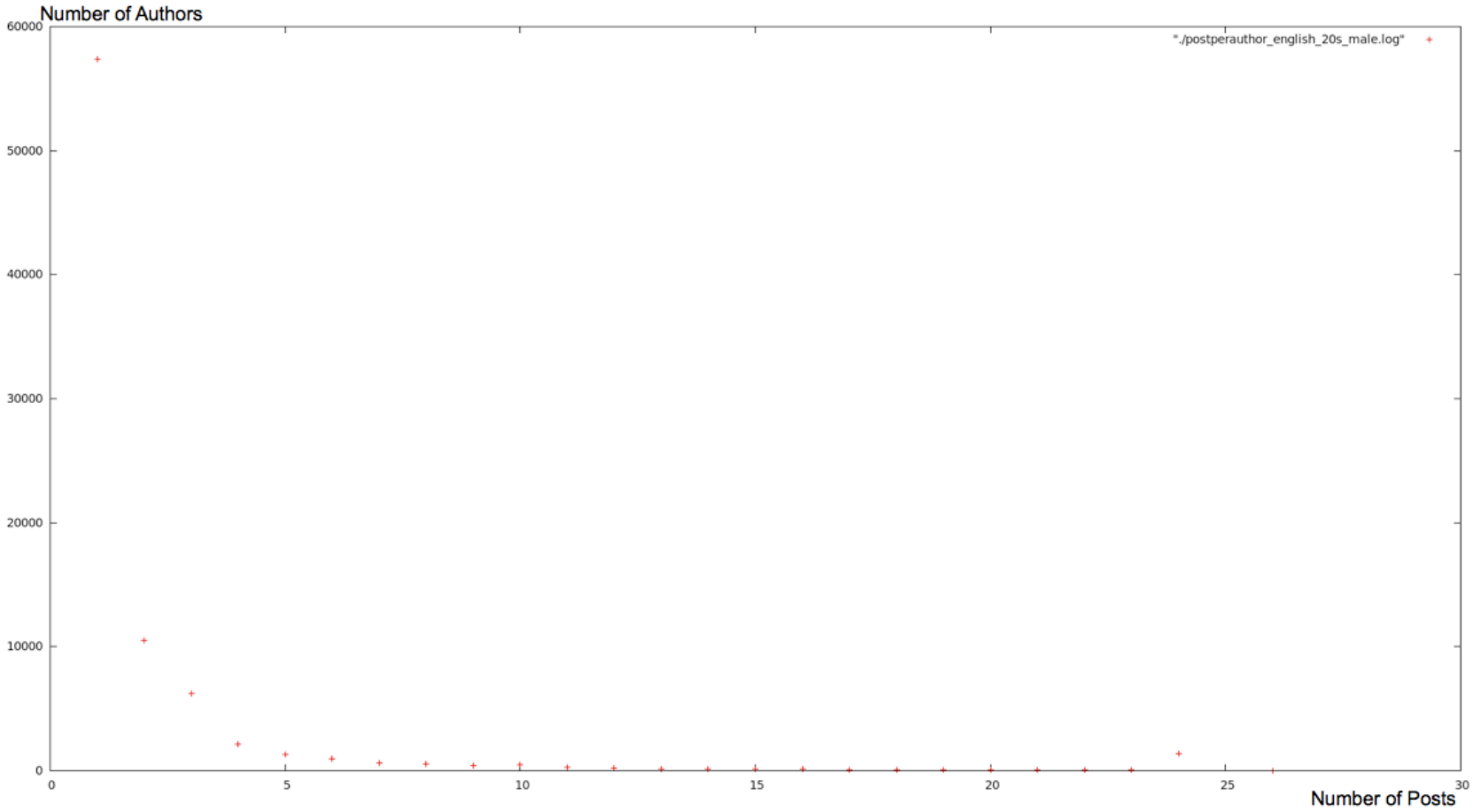
P.p.a.: English 10s female



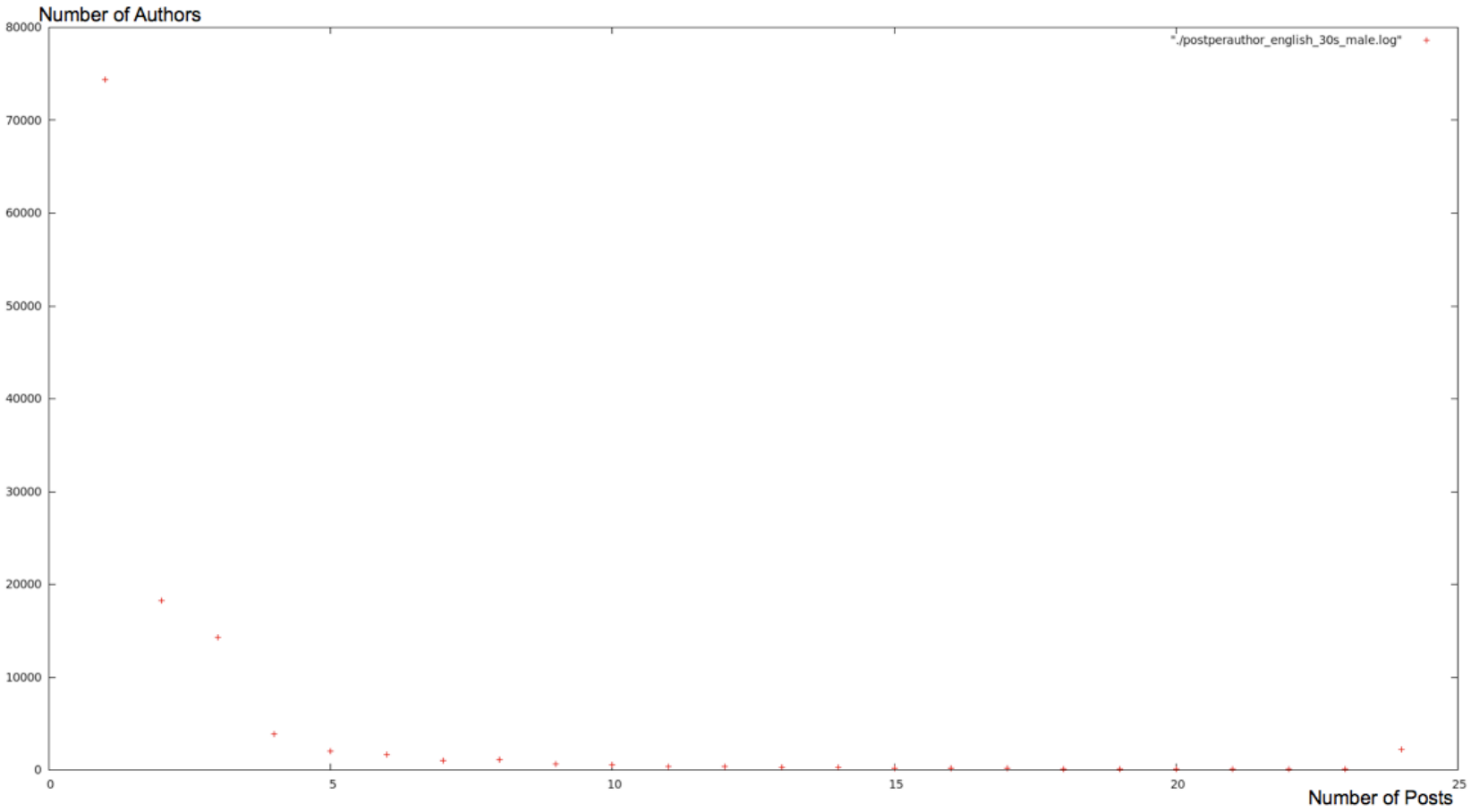
P.p.a.: English 20s male



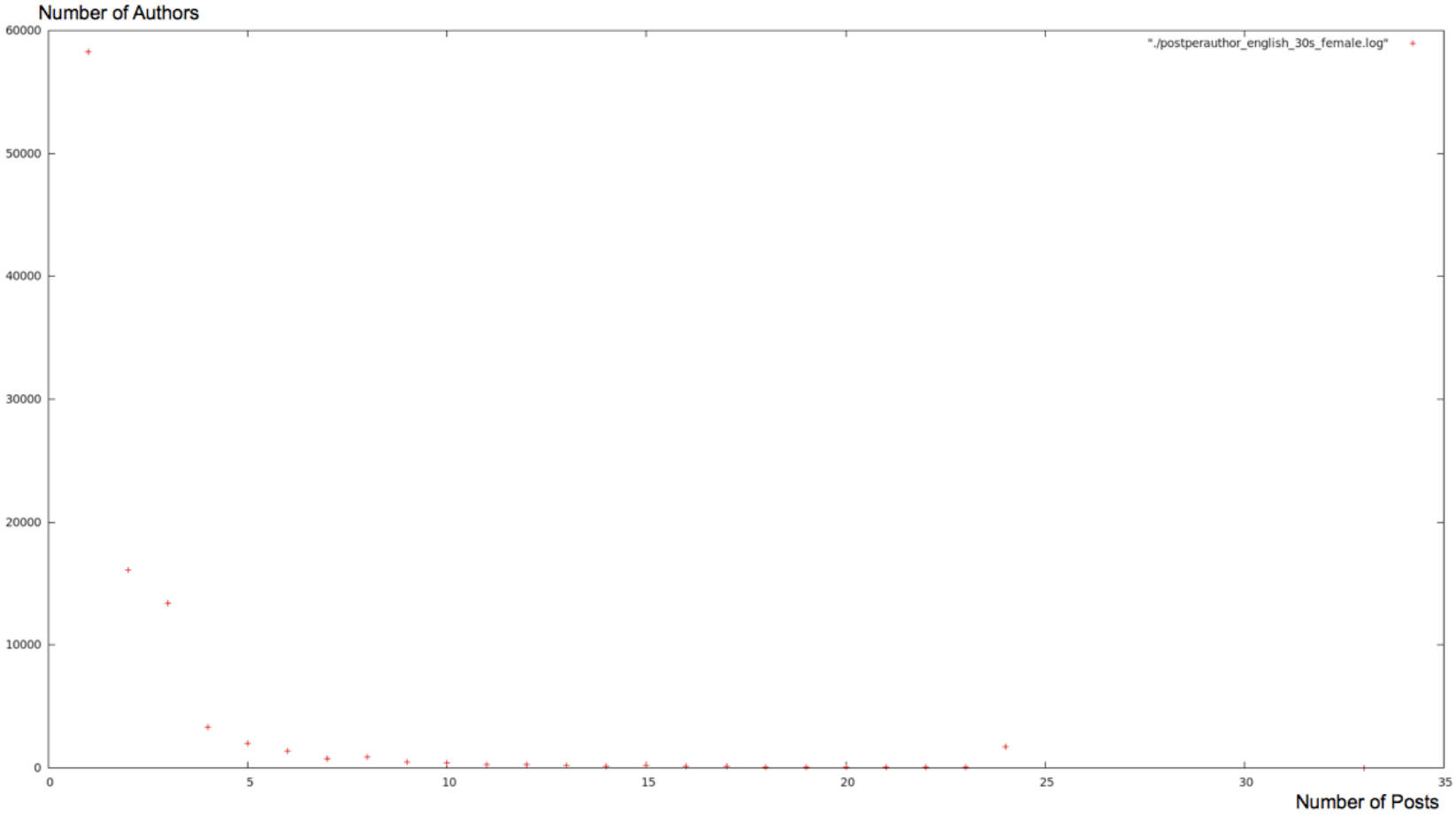
P.p.a.: English 20s female



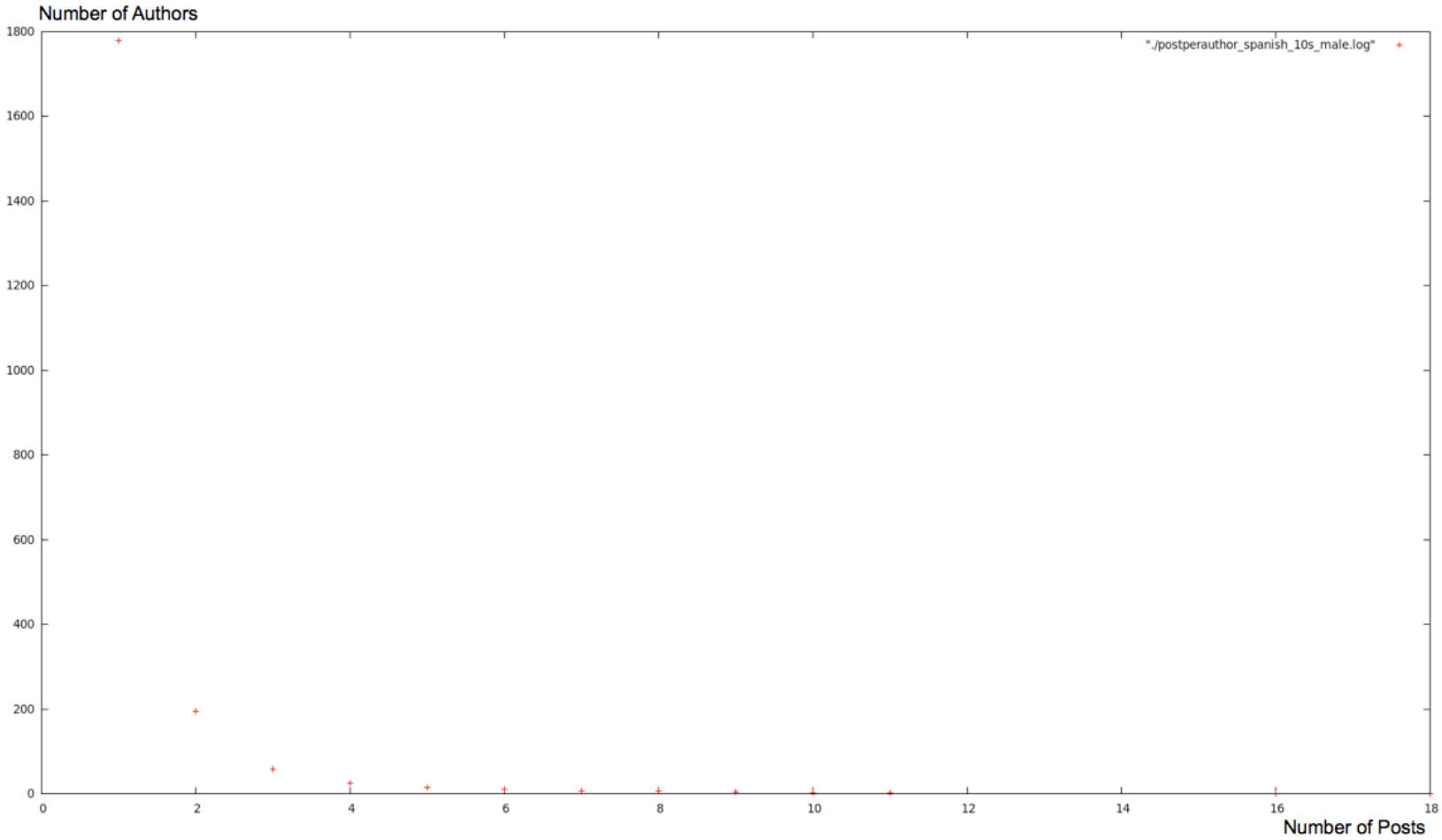
P.p.a.: English 30s male



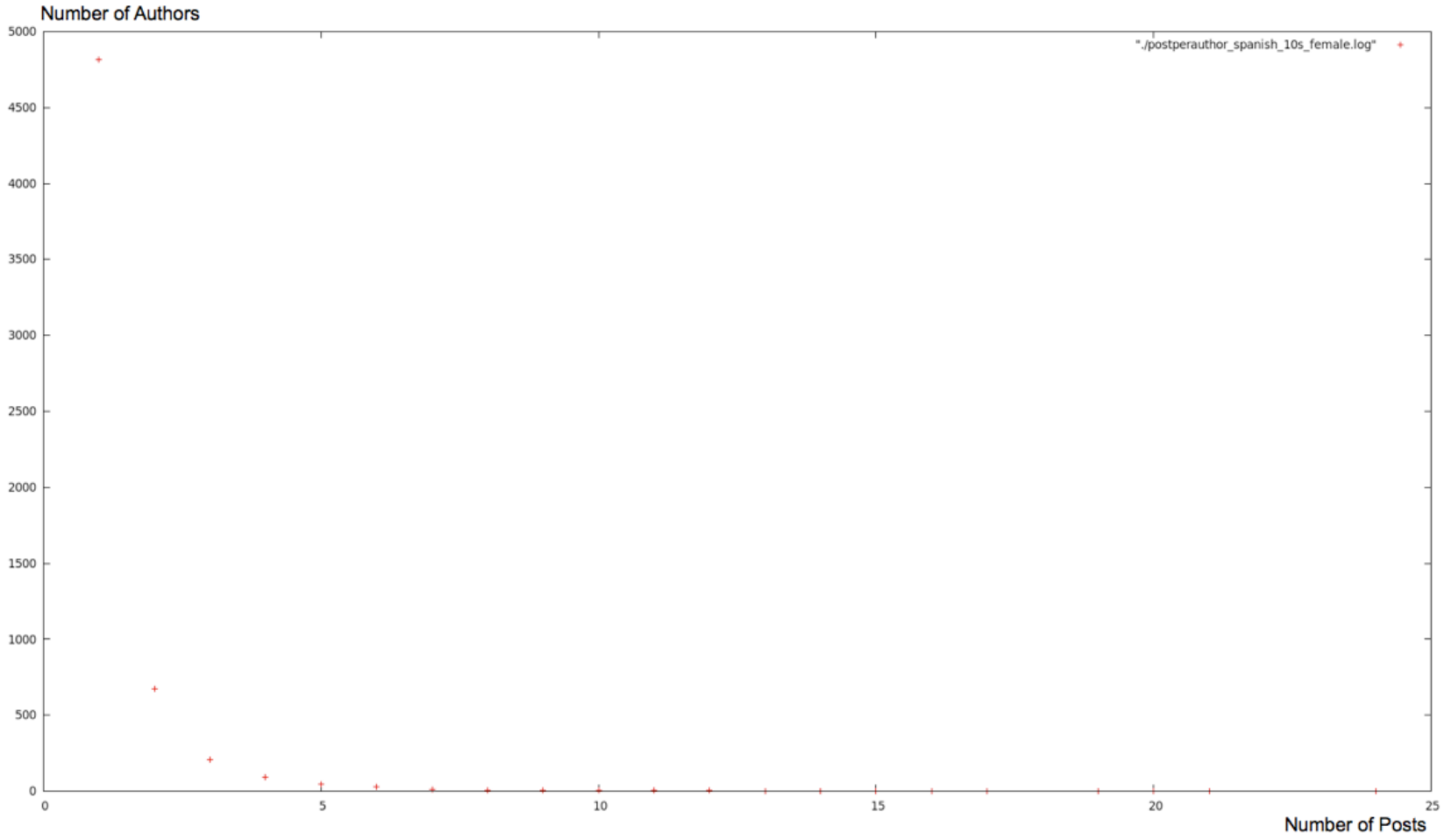
P.p.a.: English 30s female



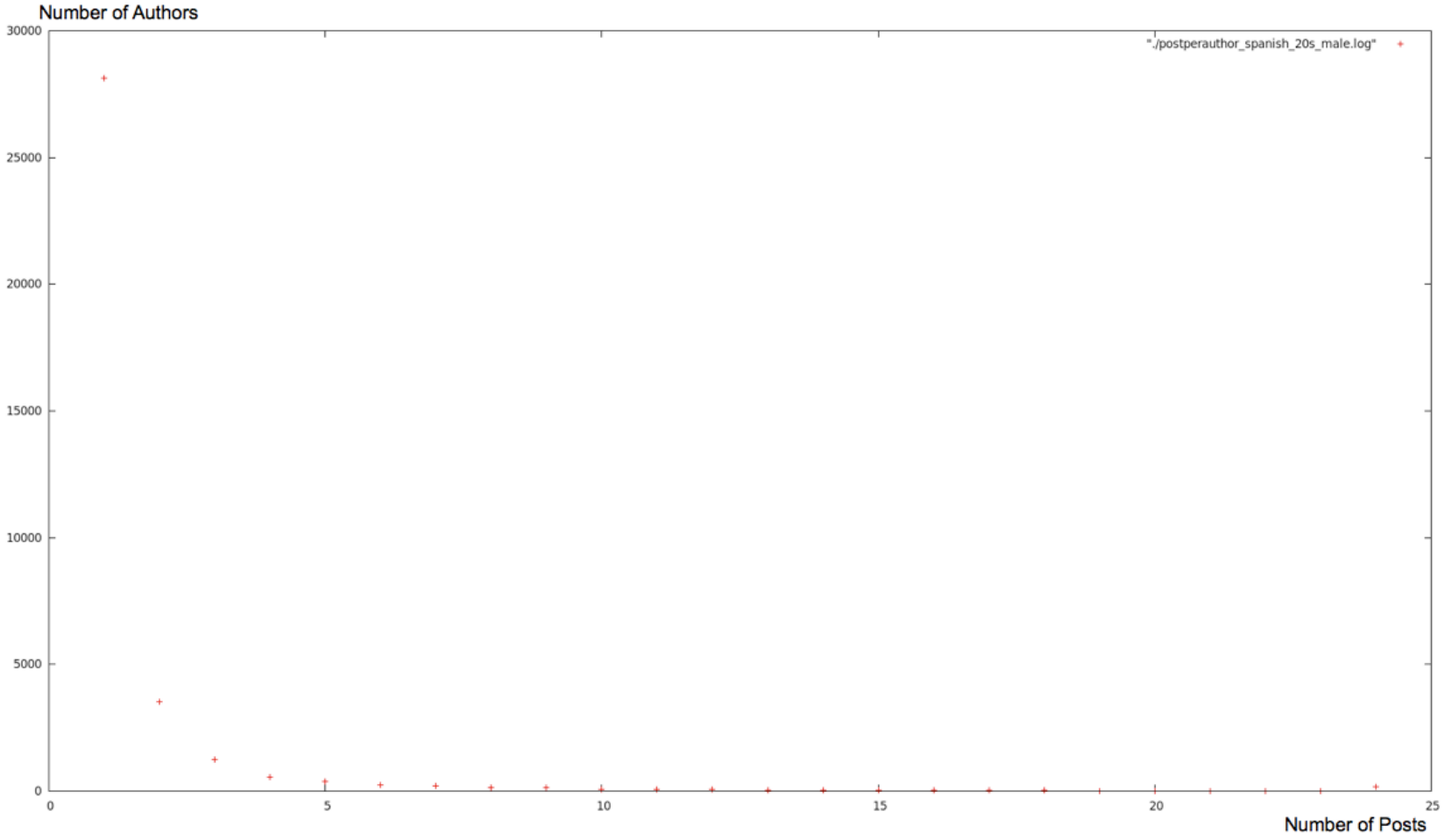
P.p.a.: Spanish 10s male



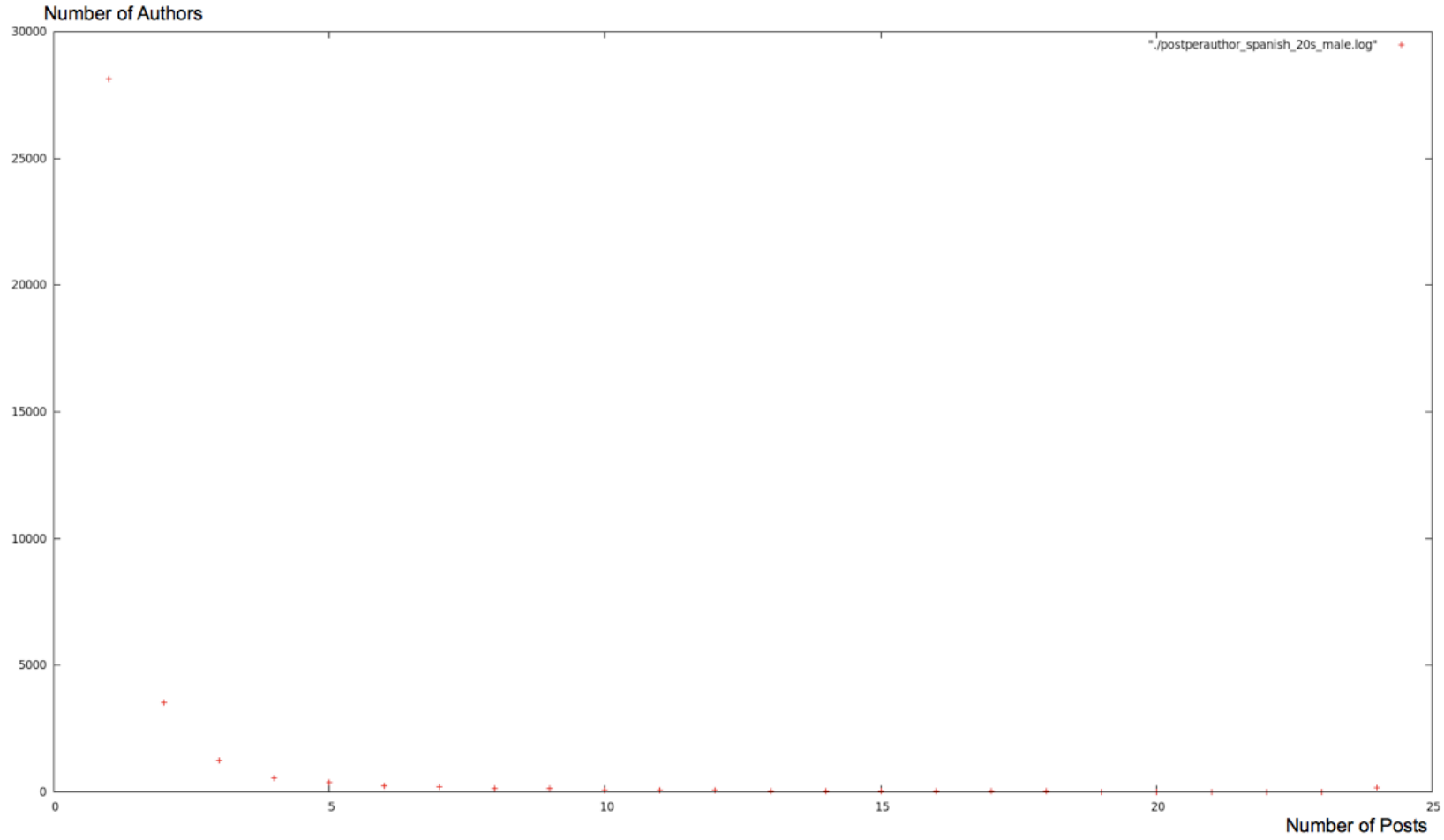
P.p.a.: Spanish 10s female



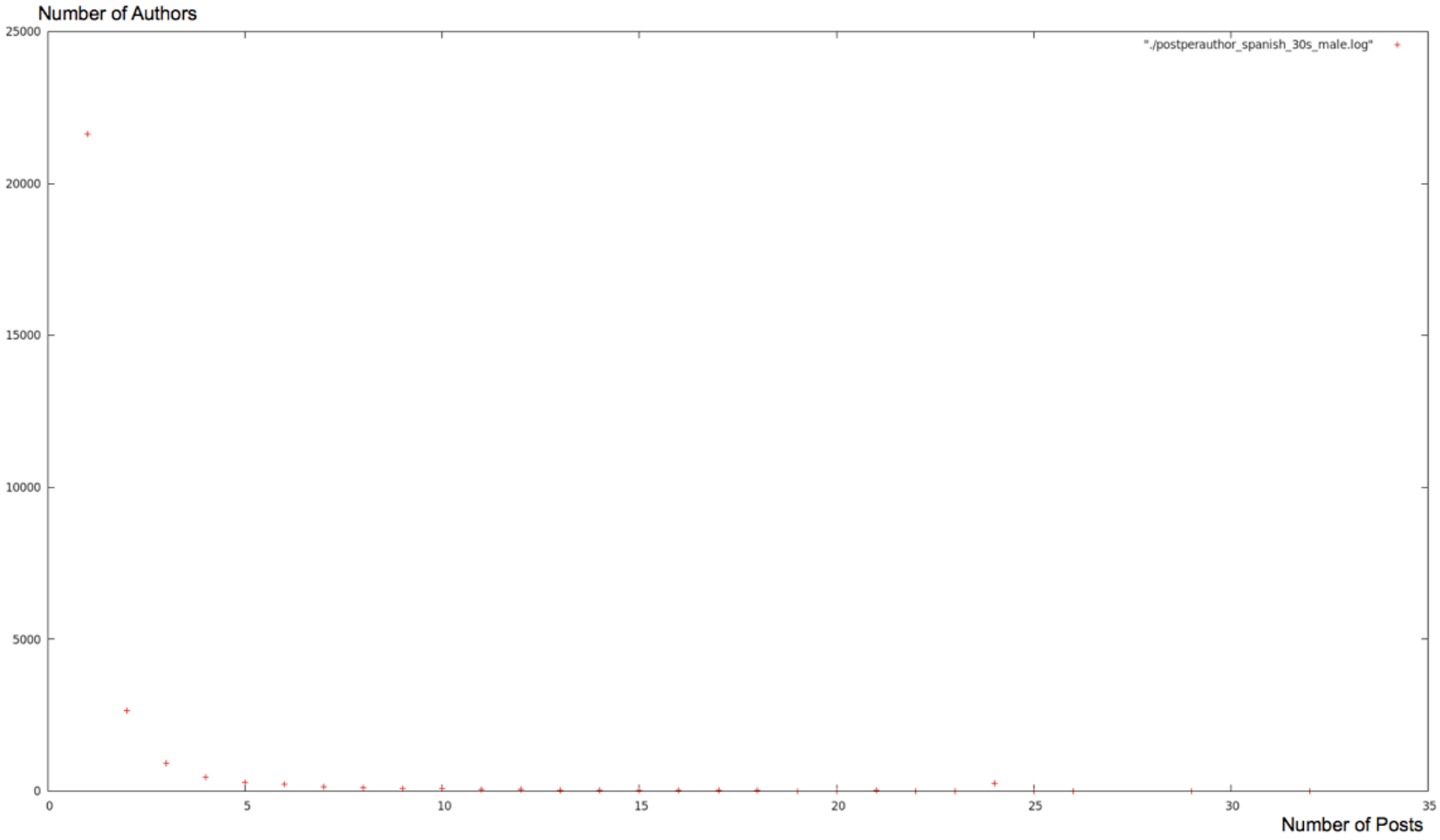
P.p.a.: Spanish 20s male



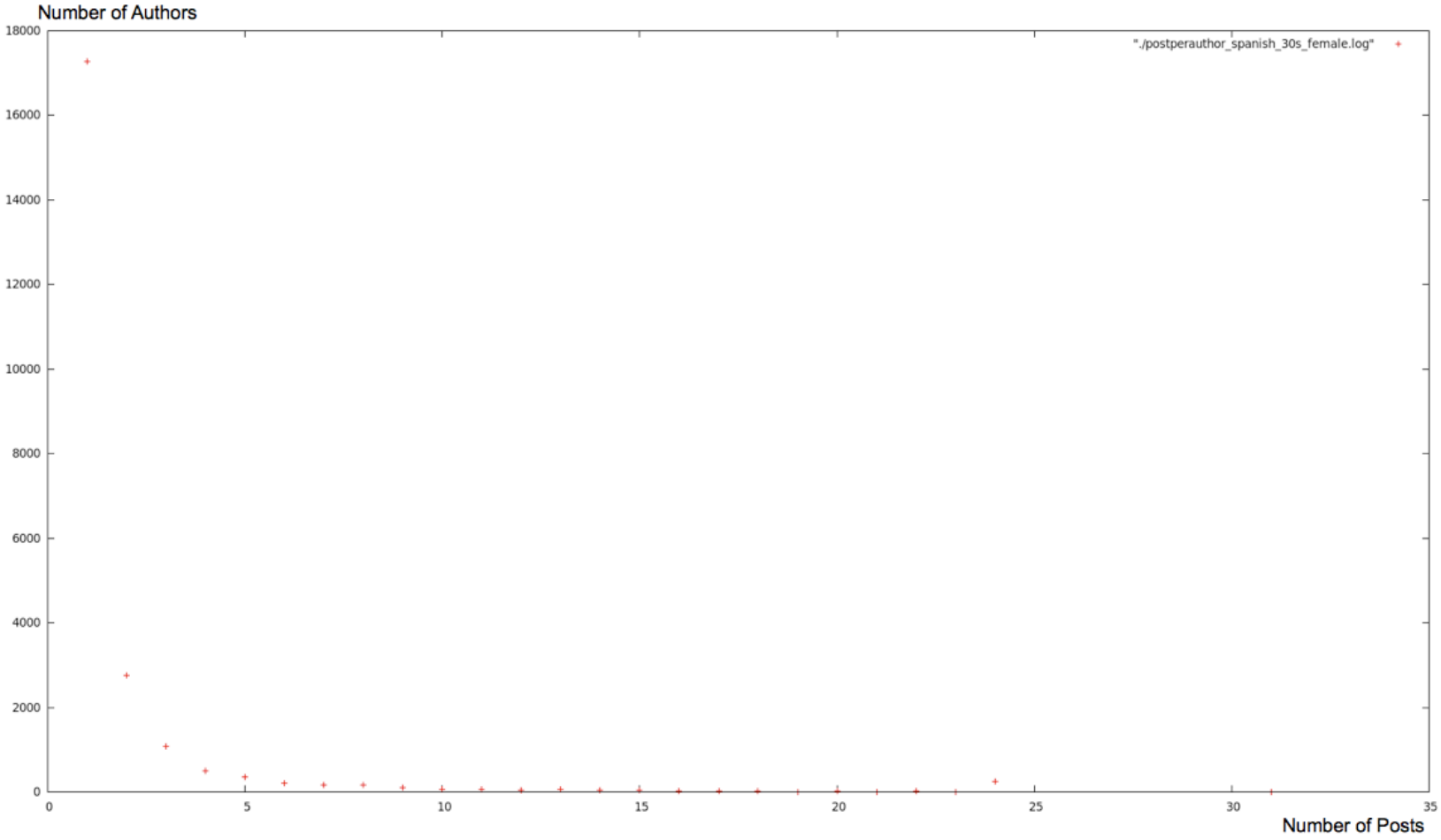
P.p.a.: Spanish 20s female



P.p.a.: Spanish 30s male



P.p.a.: Spanish 30s female



Words per author: Distribution

N. WORDS	ENGLISH						SPANISH					
	10s		20s		30s		10s		20s		30s	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
[0, 200[	1,791	3,157	43,815	30,639	29,553	16,613	1,781	4,857	27,662	24,826	21,425	18,138
[200, inf.[	10,851	11,287	39,853	35,897	93,080	84,199	330	1,077	7,644	8,958	5,886	5,176

N. WORDS	ENGLISH						SPANISH					
	10s		20s		30s		10s		20s		30s	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
[0, 200[	1,791	3,157	43,815	30,639	29,553	16,613	1,781	4,857	27,662	24,826	21,425	18,138
[200, 500[	4,836	5,141	18,848	16,746	41,437	37,670	169	645	3,789	4,930	2,540	2,347
[500, inf.[	6,015	6,146	21,005	19,151	51,643	46,529	161	432	3,855	4,028	3,346	2,829

N. WORDS	ENGLISH						SPANISH					
	10s		20s		30s		10s		20s		30s	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
[0, 50[	1,245	1,814	29,975	18,474	19,213	9,919	1,237	2,809	17,163	13,533	14,571	11,698
[50, 100[	299	714	8,163	6,996	6,009	3,820	348	1,194	6,370	6,607	4,179	3,790
[100, 300[	698	1,192	9,642	8,558	9,222	6,777	278	1,166	6,029	7,162	3,890	3,805
[300, 1000[	6,916	7,255	24,086	21,542	58,621	53,702	171	597	3,667	4,521	2,699	2,422
[1000, inf.[	3,482	3,469	11,802	10,966	29,568	26,594	77	168	2,077	1,961	1,972	1,599

Data Selection Criteria

- Group by author
- Keep authors with few post
- Chunk authors with more than 1,000 words
- Random split in three datasets
  - Training
  - Early Bird
  - Testing
- Balance number of authors per gender
- Introduce few special cases
  - Predators
  - Adult-adult sexual conversations
- Balance by gender
- Age groups:
  - 10s (13-17)
  - 20s (23-27)
  - 30s (33-47)

Train distribution

LANG	AGE	GENDER	NUM. AUTHORS TRAIN	NUM. AUTHORS TRAIN	NUM. AUTHORS TRAIN
EN	10s	MALE	8,600	17,200	236,600
		FEMALE	8,600		
	20s	MALE	42,828 72	85,800	
		FEMALE	42,875 25		
	30s	MALE	66,708 92	133,600	
		FEMALE	66,800		
ES	10s	MALE	1,250	2,500	75,900
		FEMALE	1,250		
	20s	MALE	21,300	42,600	
		FEMALE	21,300		
	30s	MALE	15,400	30,800	
		FEMALE	15,400		

Legend

123: Data
123: Pedophiles Data
123: Adult-adult sexual chats

Early/+20% distribution

LANG	AGE	GENDER	NUM. AUTHORS EARLY / +20%	NUM. AUTHORS EARLY / +20%	NUM. AUTHORS EARLY / +20%
EN	10s	MALE	740 / 148	1,480 / + 296	21,200 / +4,240
		FEMALE	740 / 148		
	20s	MALE	3,840 / 736 0 / 32	7,680 / +1,536	
		FEMALE	3,840 / 758 0 / 10		
	30s	MALE	6,020 / 1,164 0 / 40	12,040 / +2,408	
		FEMALE	6,020 / +1,204		
ES	10s	MALE	120 / +24	240 / +48	6,800 / +1,360
		FEMALE	120 / +24		
	20s	MALE	1,920 / +384	3,840 / +768	
		FEMALE	1,920 / +384		
	30s	MALE	1,360 / +272	2,720 / +544	
		FEMALE	1,360 / +272		
				Legend	123: Data 123: Pedophiles Data 123: Adult-adult sexual chats

Early bird results

TEAM	ENGLISH		SPANISH	
	AGE	GENDER	AGE	GENDER
GILLAN	**59.47**	54.13	53.57	47.74
LANDRA	59.24	**56.31**	57.57	**61.71**
AYALA	2.78	2.77	8.41	8.44
JANKOWSKA	54.63	51.84	44.79	58.34
BASELINE	33.24	49.97	33.53	50.01

Final results

TEAM	ENGLISH			SPANISH				
	AGE	GENDER	BOTH	AGE	GENDER	BOTH	TOTAL	TIME
P1	**65.72**	**56.90**	**38.13**	**65.58**	62.99	41.58	**39.86**	38.31m
P2	64.08	56.52	35.08	64.30	**64.73**	**42.08**	38.58	4.86h
P3	59.66	54.56	31.14	62.19	61.65	38.97	35.06	2,66h
P4	61.18	56.08	34.20	57.27	61.38	35.23	34.72	28.83m
P5	60.98	56.71	34.88	57.05	54.68	31.20	33.04	6.69d
P6	56.90	52.67	31.15	53.75	57.06	31.34	31.25	2.12d
P7	60.31	54.10	32.68	53.77	47.84	25.43	29.06	**10.26m**
P8	60.55	50.00	28.43	56.43	50.00	28.24	28.34	1.04h
P9	54.15	47.81	24.71	56.51	51.16	29.34	27.03	17.88h
P10	47.38	53.81	28.14	42.76	58.46	25.92	27.03	4,66h
P11	59.23	55.22	32.92	0.00	0.00	0.00	16.46	2.87d
P12	37.53	37.88	19.86	0.00	0.00	0.00	9.93	2.77h

- There are still 8 pending evaluations

Preliminary conclusions

- Very difficult task, mainly for gender detection
- Difficult to identify together both gender and age
- Expensive in Time consuming

Experiments with Author Profiling dataset PAN 2013

Comparison with Early Birds

TEAM	ENGLISH		SPANISH	
	AGE	GENDER	AGE	GENDER
GILLAN	**59.47**	54.13	53.57	47.74
LANDRA	59.24	**56.31**	57.57	**61.71**
AYALA	2.78	2.77	8.41	8.44
JANKOWSKA	54.63	51.84	44.79	58.34
BASELINE	33.24	49.97	33.53	50.01
RANGEL	-	-	**62.72**	56.75

Do not forget, we are watching you! ;-)