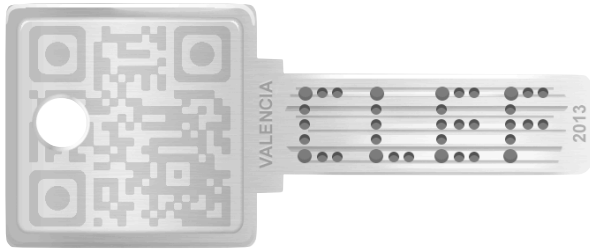
Author Profiling

PAN-AP-2013 - CLEF 2013

Valencia, 24th September 2013

PAN



Francisco Rangel
Autoritas / Universitat Politècnica de València

Paolo Rosso
Universitat Politècnica de València

Moshe Koppel
Bar-Illan University

Efstathios Stamatatos
University of the Aegean

Giacomo Inches
University of Lugano

What's Author Profiling?

Gender?

Personality traits?

Age?

Emotions?

Native language?

Author Profile... Who is who?

Why Author Profiling?

Forensics	Security	Marketing
Language as evidence	***Profile possible delinquents***	***Segmenting users***

Task Goals

- Given a collection of documents retrieved from Social Media in English and Spanish...

MAIN GOAL	SECONDARY GOALS
Identify age and gender	Test the robustness of the approaches for identifying age and gender of predators
	Measure the computational time needed to perform the task

Related Work

AUTHOR	COLLECTION	FEATURES	RESULTS	OTHER CHARACTERISTICS
Argamon et al., 2002	British National Corpus	Part-of-speech	Gender: 80% accuracy	
Holmes & Meyerhoff, 2003	Formal texts	-	Age and gender	
Burger & Henderson, 2006	Blogs	Posts length, capital letters, punctuations. HTML features.	They only reported: "Low percentage errors"	Two age classes: [0,18[,[18,-]
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	Blogs	Slang + sentence length	Gender: 89.18 accuracy Age: 80.32 accuracy	
Zhang & Zhang, 2010	Segments of blog	Words, punctuation, average words/sentence length, POS, word factor analysis	Gender: 72,10 accuracy	
Nguyen et al., 2011 y 2013	Blogs & Twitter	Unigrams, POS, LIWC	Correlation: 0.74 Mean absolute error: 4.1 - 6.8 years	Manual labeling Age as continuous variable
Peersman et al., 2011	Netlog	Unigrams, bigrams, trigrams and tetagrams	Gender+Age: 88.8 accuracy	Self-labeling, min 16 plus 16,18,25

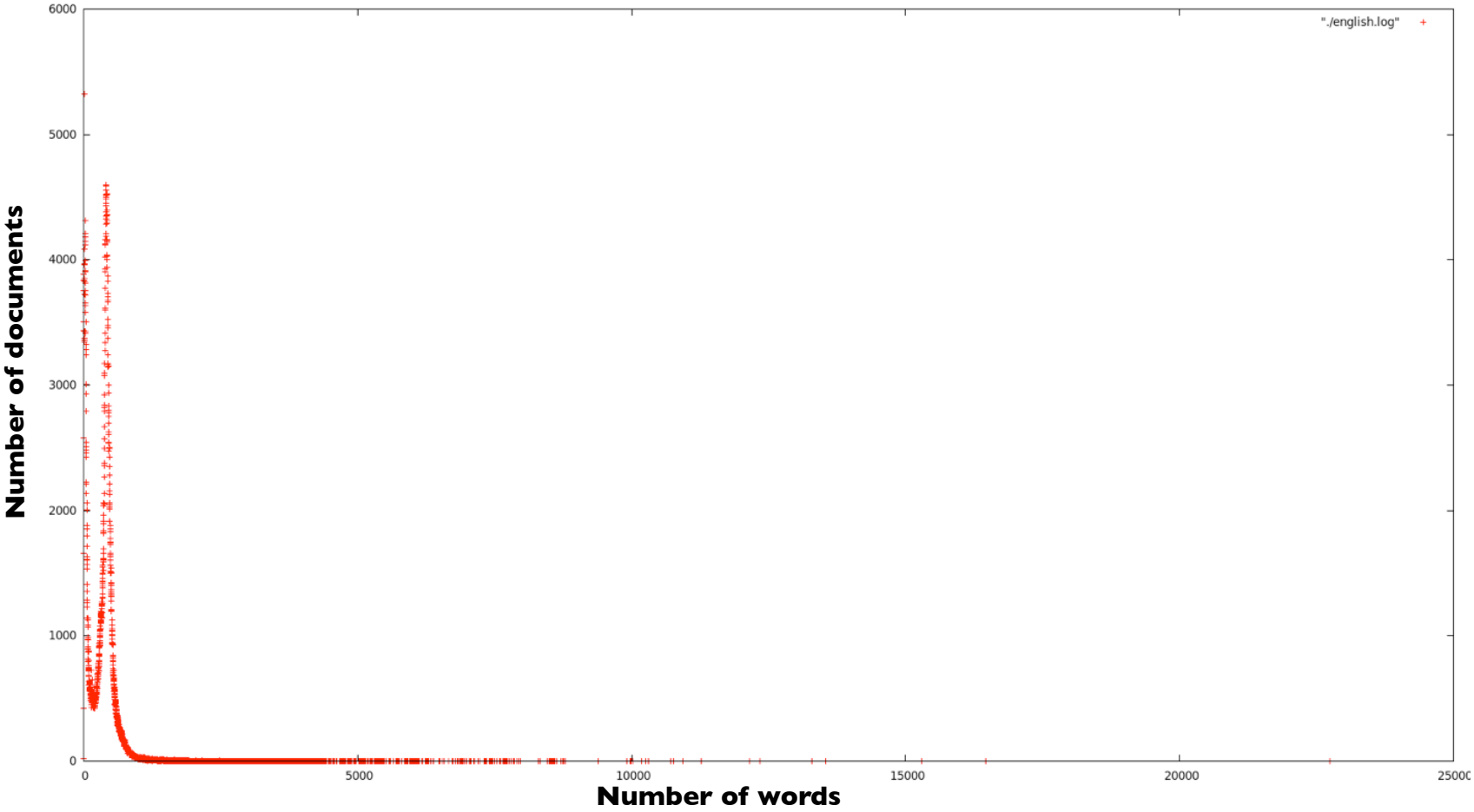
Data Collection – Social Media

- Big Data?
- High variety of themes
- Sexual conversations vs. sexual predators
- Difficulty to obtain good label data
- Real people vs. Robots (chatbots)
- Multilingual: English + Spanish



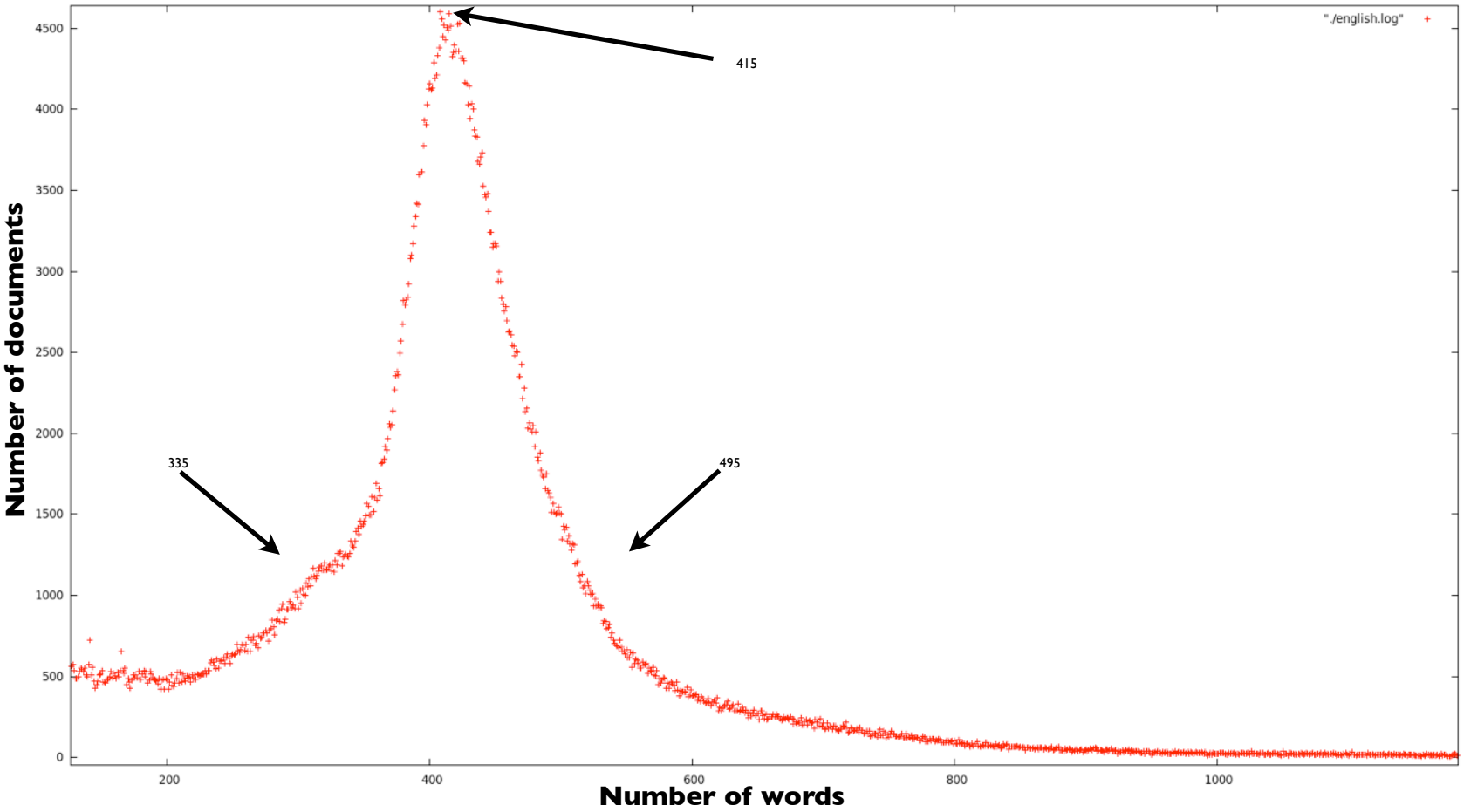



Data Collection – English Distribution



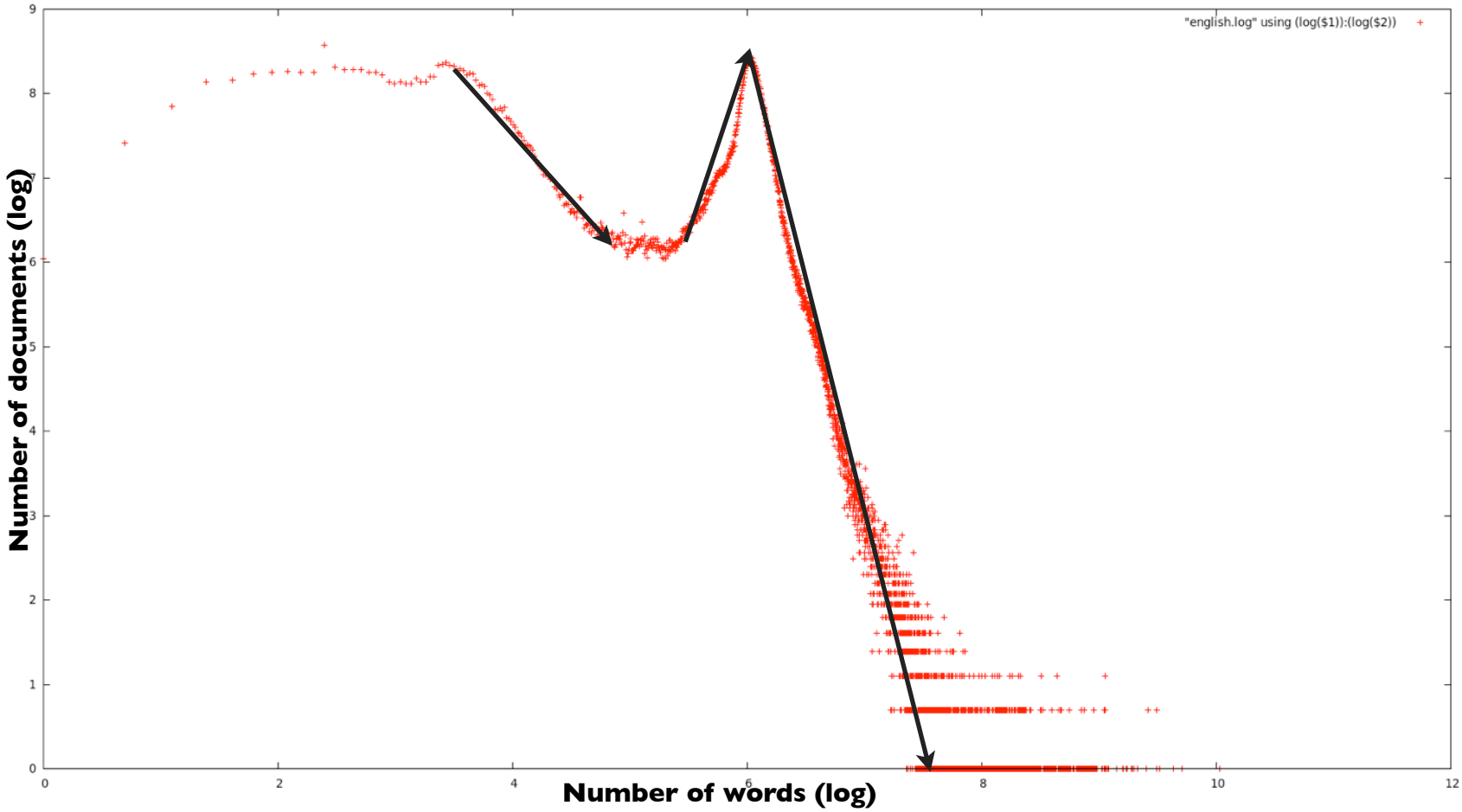
MIN	MAX	AVG	STD
0	22,736	335	208

Data Collection – English Distribution (zoomed)



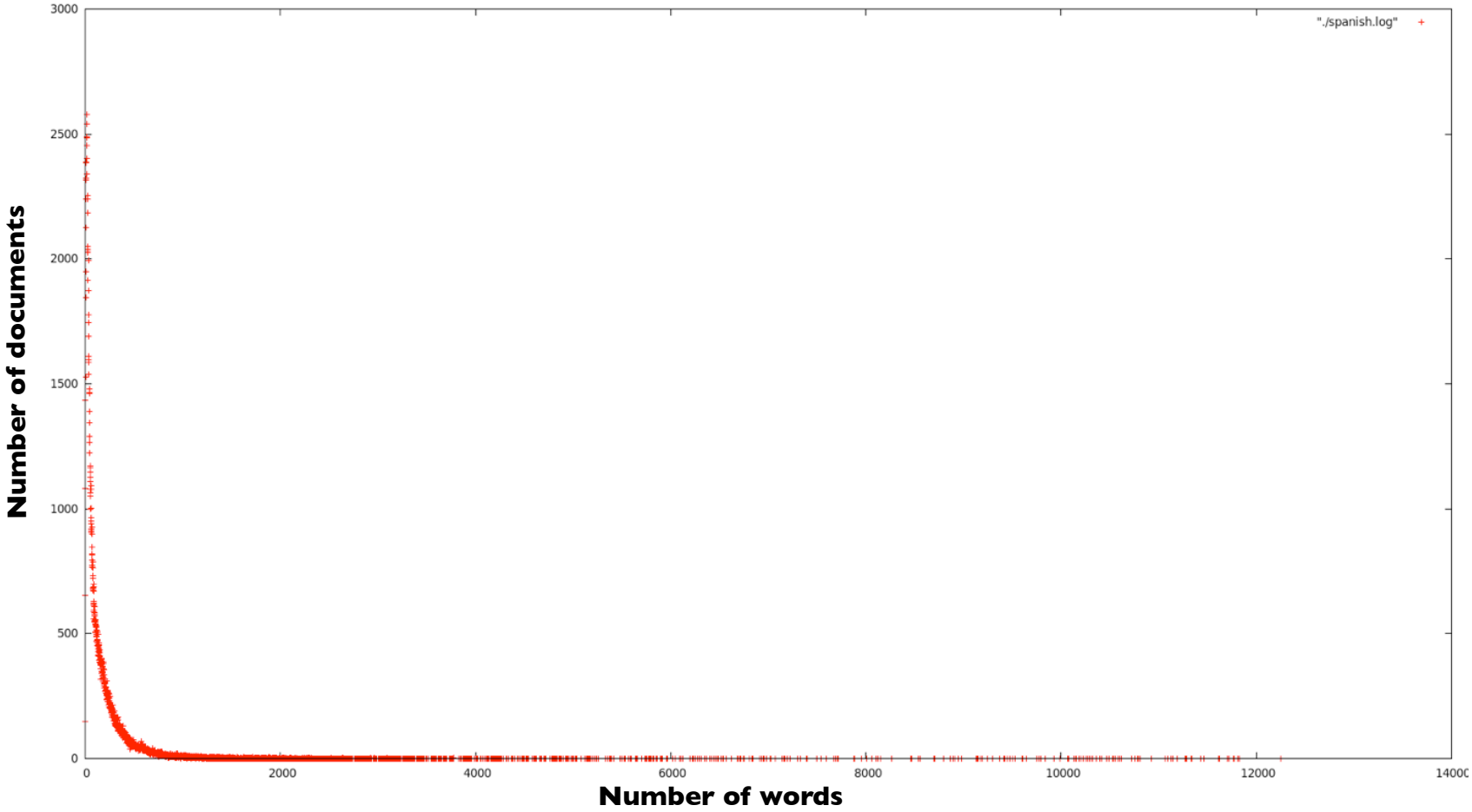
- If we zoom the distribution, we can observe a gaussian like distribution, with its maximum on the value 415.

Data Collection – English Distribution (log-log)



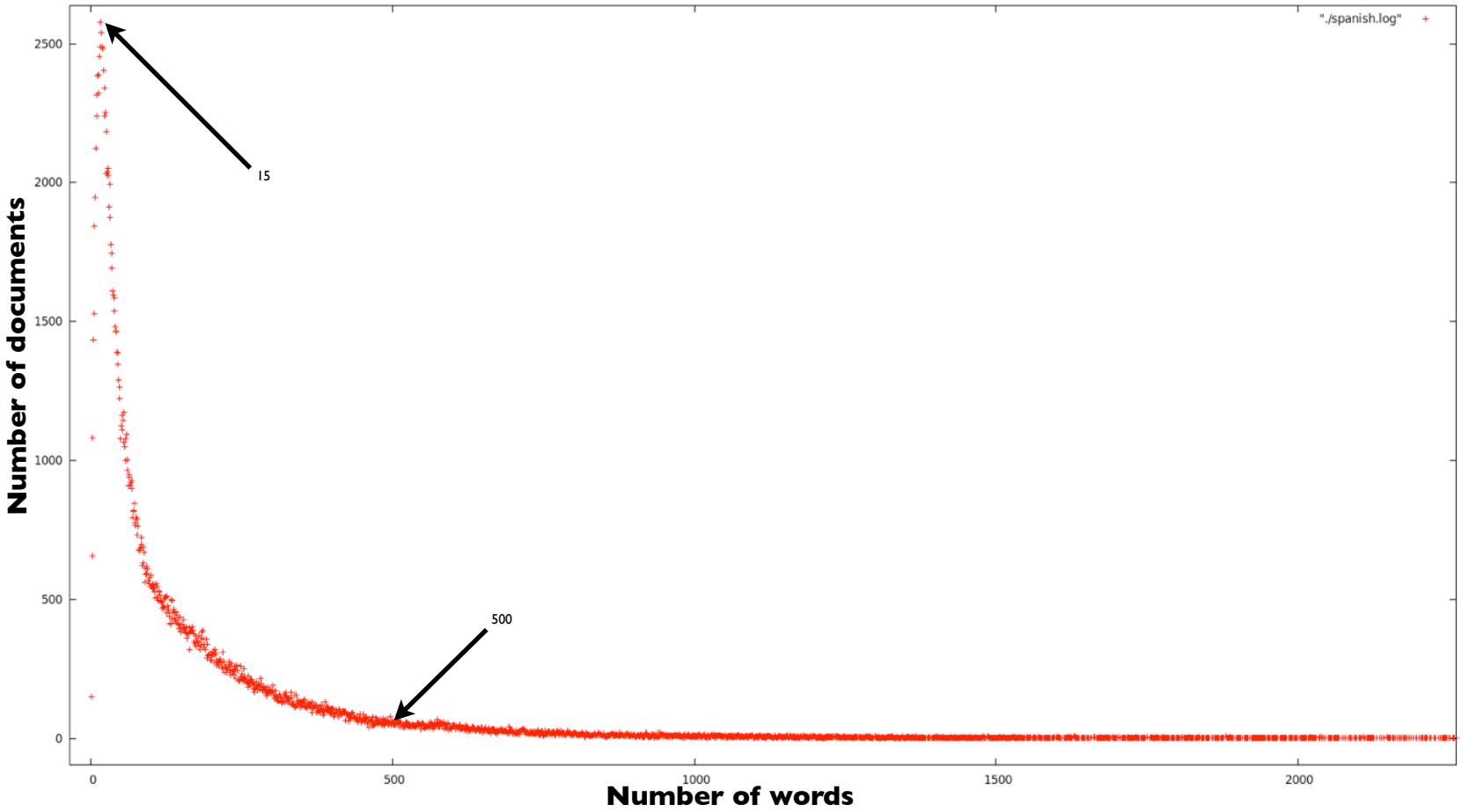
- The log-log representation shows how the distribution has a long tail component, specifically in two cases, one before the point of maximum frequency and another one after this.
- We could use this property to select minimum and maximum number of words that the posts must have.

Data Collection – Spanish Distribution

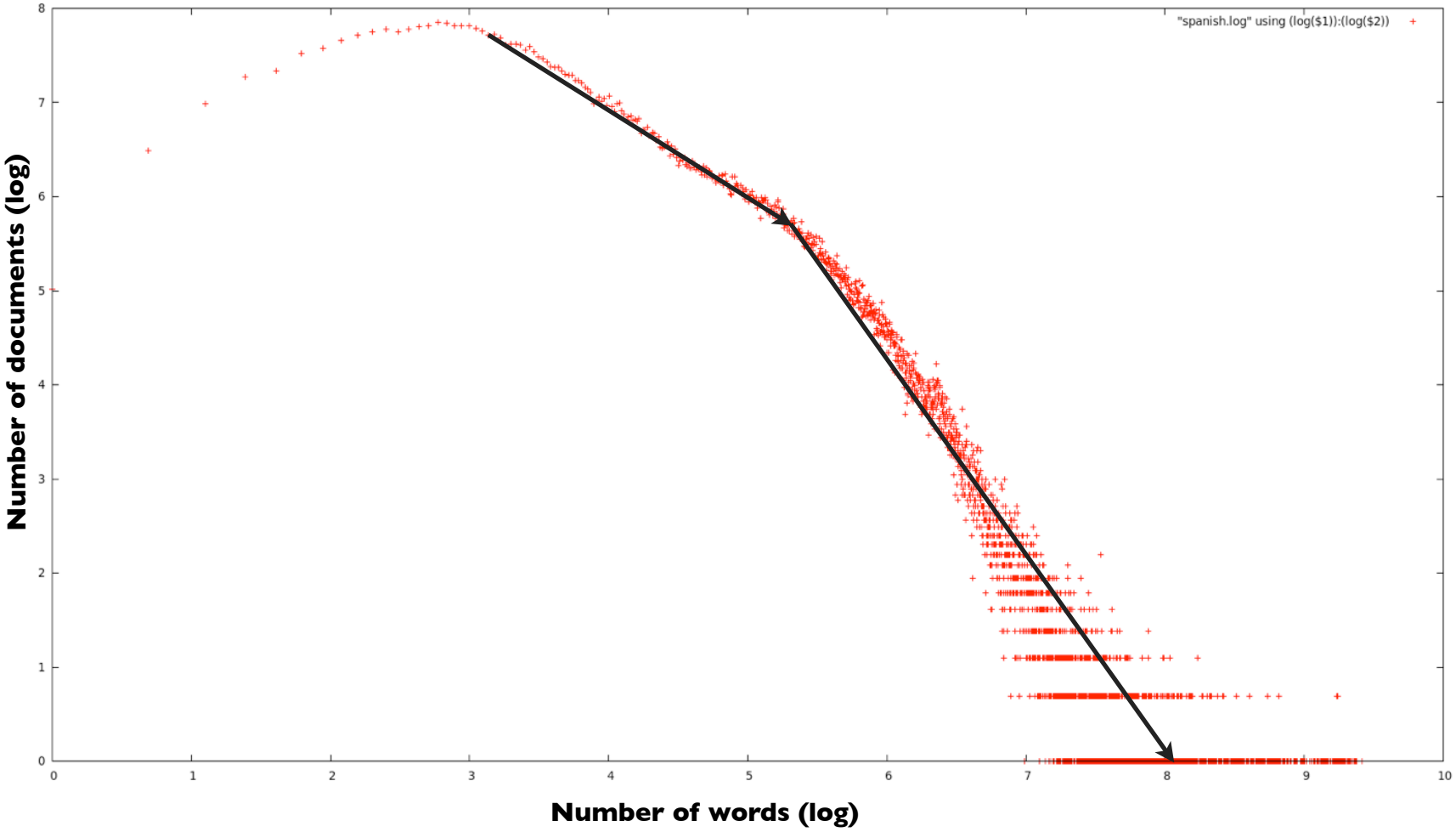


MIN	MAX	AVG	STD
0	12,246	176	832

Data Collection – Spanish Distribution (zoomed)



Data Collection – Spanish Distribution (log-log)



Data Collection – Selection Criteria

- Grouping posts by author
- Keeping authors with few post
- Chunking authors with more than 1,000 words

- Balanced by gender
- Age groups (non-balanced):
  - 10s (13-17)
  - 20s (23-27)
  - 30s (33-47)

- Random split in three datasets
  - Training
  - Early Bird (10%)
  - Testing (+20%)

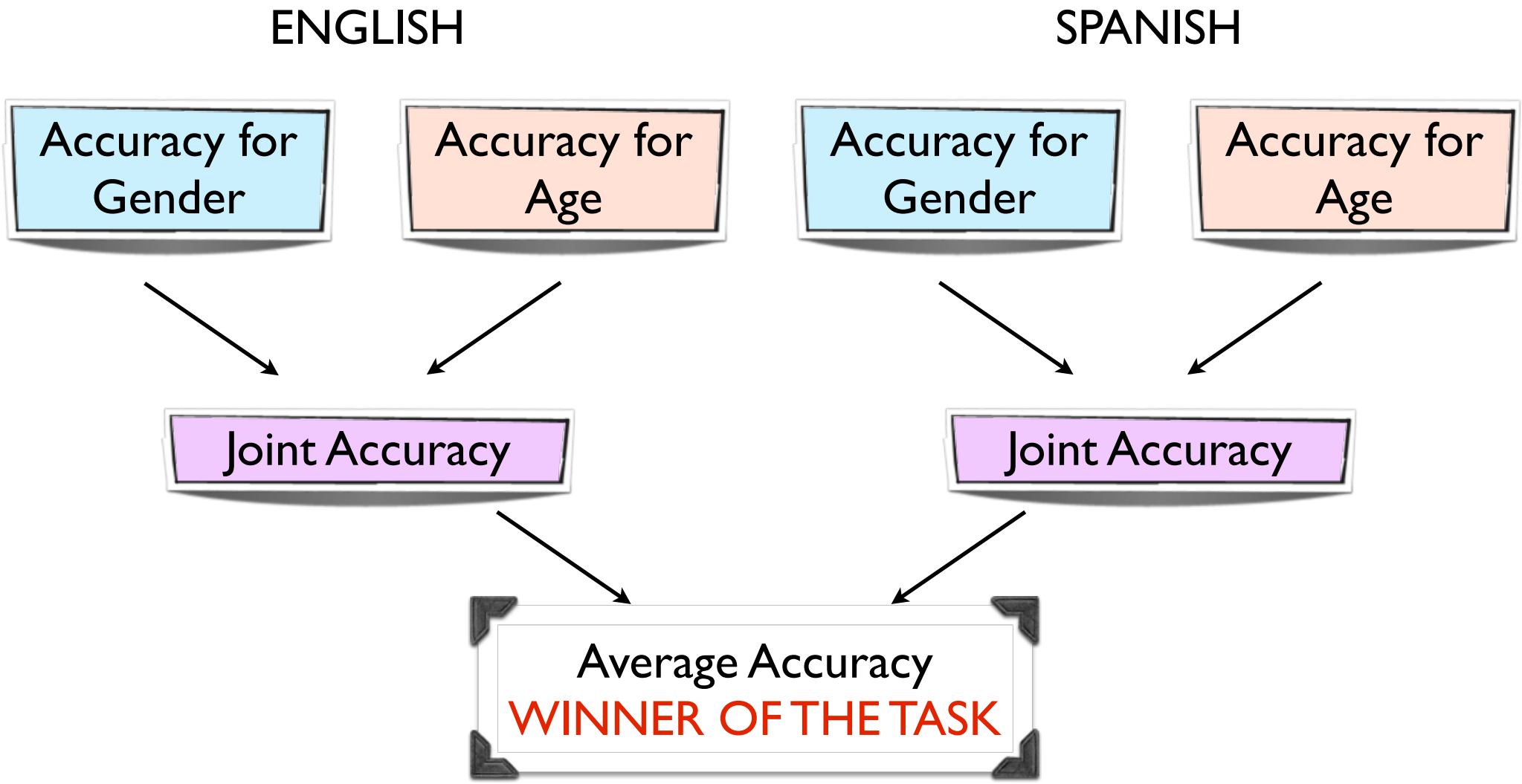
- Introduction of few special cases
  - Predators (0.0012%)
  - Adult-adult sexual conversations

Data Collection – Statistics

LANG AGE GENDER			NUMBER OF AUTHORS		
			TRAINING	EARLY BIRDS	TEST
EN	10s	MALE	8,600	740	888
		FEMALE	8,600	740	888
	20s	MALE	(72) 42,828	3,840	(32) 4,576
		FEMALE	(25) 42,875	3,840	(10) 4,576
	30s	MALE	(92) 66,708	6,020	(40) 7,184
		FEMALE	66,800	6,020	7,224
∑			***236,600***	***21,200***	***25,440***
ES	10s	MALE	1,250	120	144
		FEMALE	1,250	120	144
	20s	MALE	21,300	1,920	2,304
		FEMALE	21,300	1,920	2,304
	30s	MALE	15,400	1,360	1,632
		FEMALE	15,400	1,360	1,632
∑			***75,900***	***6,800***	***8,160***

Predators
Adult-adult sexual conversations

Performance measures for identification



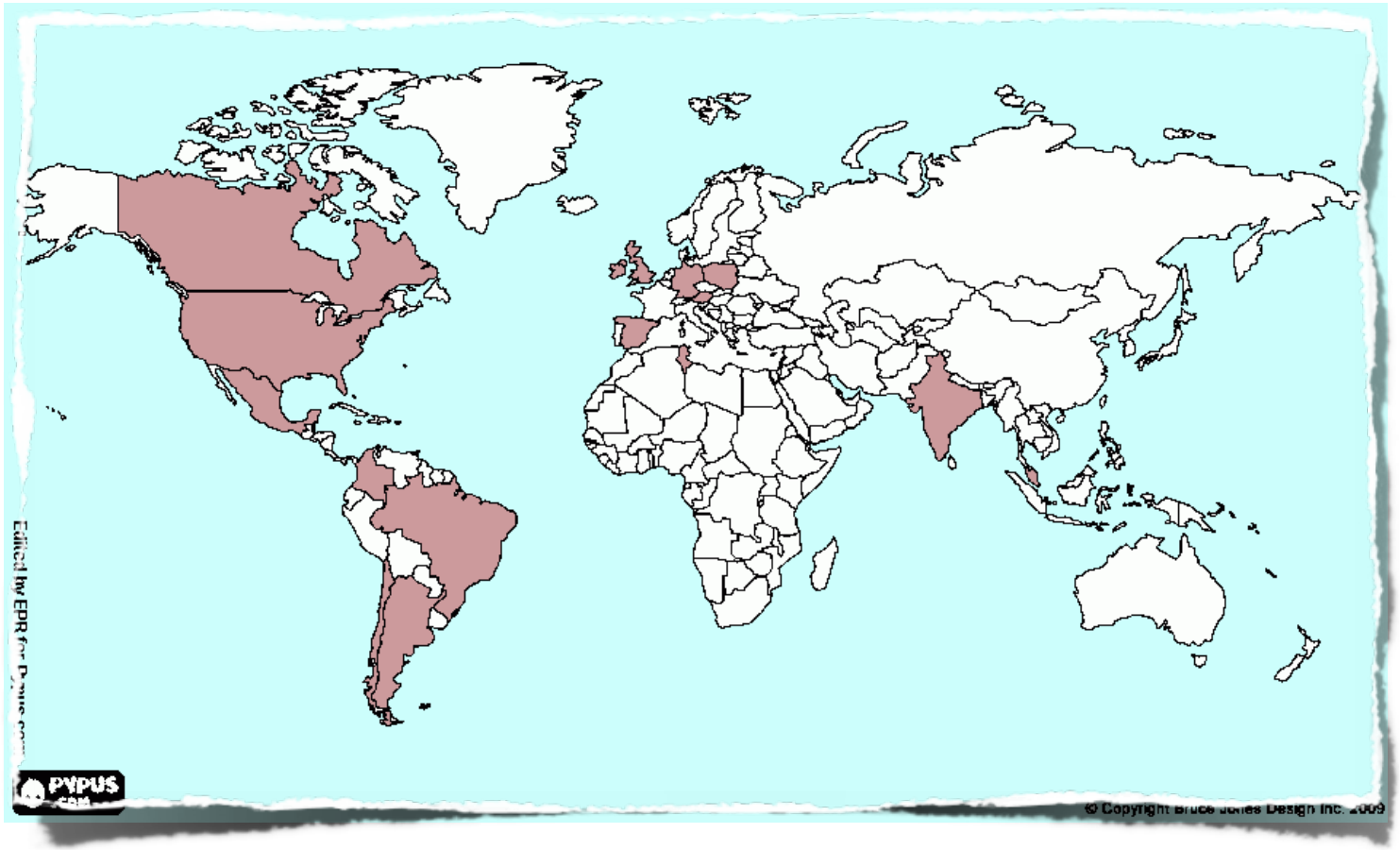
Other performance measures

Number of correctly identified gender and age for sexual conversations between adults

Number of correctly identified gender and age for predators

Total time needed to process the test data

Participants



- 66 registered teams
- 21 participants (32%)
- 16 countries
- 18 papers (86%)
  - 8 long papers
  - 10 short papers

Approaches

- What kind of ...

Preprocessing

Features

Methods

... did the teams perform?

Approaches

Preprocessing

HTML Cleaning to obtain plain text	5 teams: [gopal-patra][moreau][meina] [weren][pavan]
Deletion of documents with at least 0.1% of spam words	1 team: [flekova]
Principal Component Analysis to reduce dimensionality	1 team: [yong-lim]
Subset selection during training to reduce dimensionality	5 teams: [caurcel-diaz][flekova][moreau] [hernandez-farias][sapkota]
Discrimination between human-like posts and spam-like posts (chatbots)	1 team: [meina]

Approaches

Features

Stylistic features: frequencies of punctuation marks, capital letters, quotations...	9 teams: [yong-lim][cruz][pavan][gopal-patra][de-arteaga][meina][flekova][aleman][santosh]
+ POS tags	5 teams: [yong lim][meina][aleman][cruz][santosh]
HTML-based features like image urls or links	3 teams: [santosh][sapkota][meina]
Readability	7 teams: [gopal-patra][yong-lim][meina][flekova][aleman][weren][gillam]
Emoticons	2 teams: [aleman][hernandez-farias] *[sapkota] explicitly discarded them

Approaches

Features

Content features: LSA, BoW, TF-IDF, dictionary-based words, topic-based words, entropy-based words...	11 teams: [sapkota][gopal-patra][yong-lim][seifeddine][caurcel-diaz][flekova][meina][cruz][santosh][pavan][hernandez-farias]
Named entities	1 team: [flekova]
Sentiment words	1 team: [gopal-patra]
Emotions words	1 team: [meina]
Slang, contractions and words with character flooding	4 teams: [flekova][caurcel-diaz][aleman][hernandez-farias]

Approaches

Features

Text to be identified is used as a query for a search engine	1 team: [weren]
Unsupervised features based on statistics	1 team: [de-arteaga]
Language models (n-grams)	4 teams: [meina][jankowska][moreau] [sapkota]
Collocations	1 team: [meina]
Second order representation based on relationships between documents and profiles	1 team: [pastor]

Approaches

Methods

Decision Trees	5 teams: [santosh][gopal-patra] [seifeddine][gillam][weren]
Support Vector Machines	3 teams: [yong-lim][cruz][sapkota]
Logistic Regression	2 teams: [de-arteaga][flekova]
Naïve Bayes	1 team: [meina]
Maximum Entropy	1 team: [pavan]
Stochastic Gradient Descent	1 team: [caurcel-diaz]
Random Forest	1 team: [aleman]
Information Retrieval	1 team: [weren]

Early birds results

Table 2. Evaluation results for early birds in terms of accuracy on English (left) and Spanish (right) texts.

English			
Team	Total	Gender	Age
Ladra	0.3301	0.5631	0.5924
Gillam	0.3245	0.5413	0.5947
Jankowska	0.2796	0.5185	0.5463
baseline	0.1649	0.4997	0.3324
Aleman	0.0162	0.0277	0.0278

Spanish			
Team	Total	Gender	Age
Ladra	0.3541	0.6171	05757
Jankowska	0.2724	0.5834	0.4479
Gillam	0.2521	0.4774	0.5357
baseline	0.1653	0.5001	0.3353
Aleman	0.0490	0.0844	0.0841

- 5 teams participated, 1 team had technical problems
- Figures for gender are very close to baseline
- Main goal -> All participants improved in the final evaluation, mainly Aleman

Final results

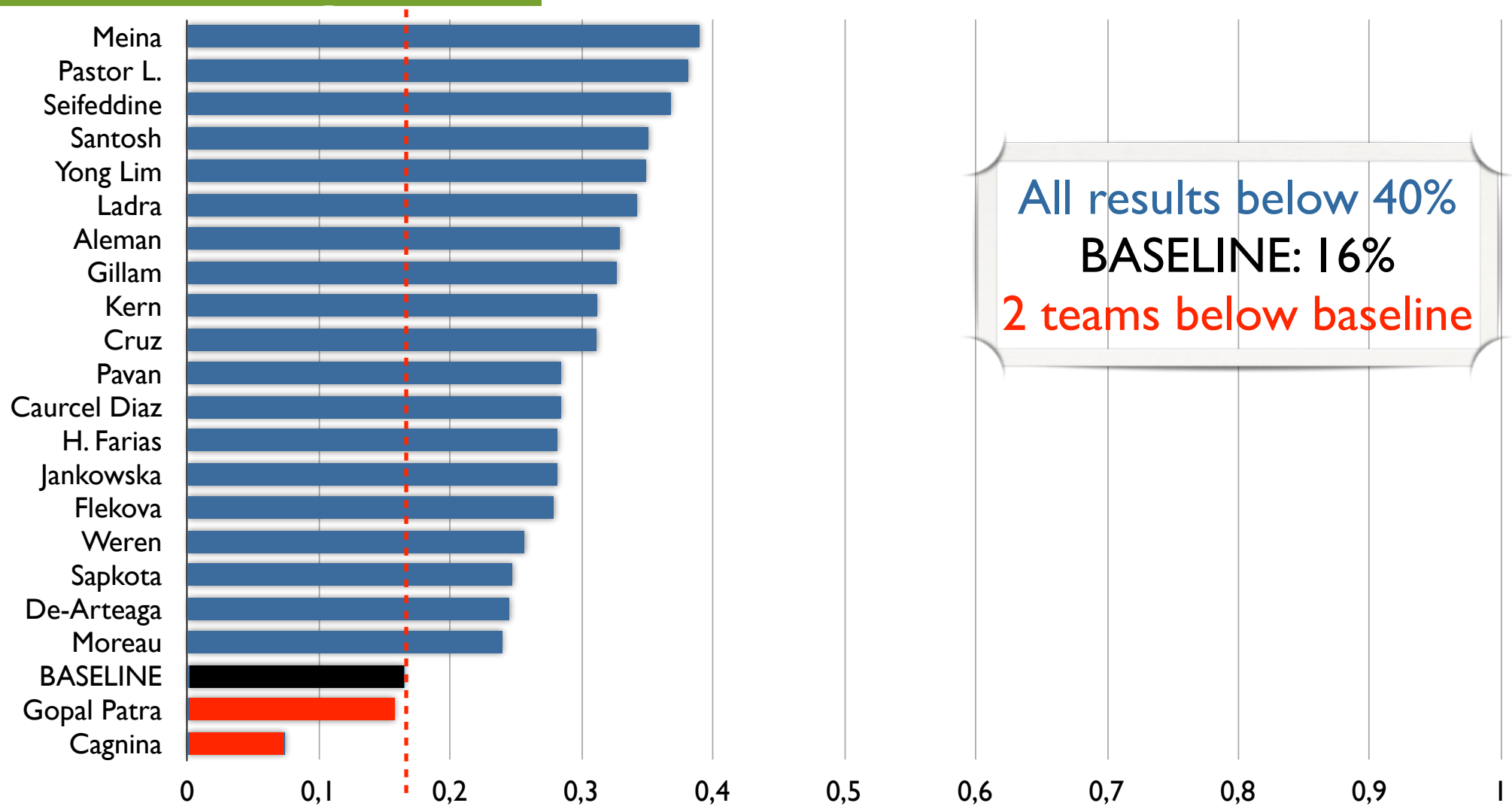
Table 3. Evaluation results in terms of accuracy on English (left) and Spansih (right) texts.

English			
Team	Total	Gender	Age
Meina	0.3894	0.5921	0.6491
Pastor L.	0.3813	0.5690	0.6572
Seifeddine	0.3677	0.5816	0.5897
Santosh	0.3508	0.5652	0.6408
Yong Lim	0.3488	0.5671	0.6098
Ladra	0.3420	0.5608	0.6118
Aleman	0.3292	0.5522	0.5923
Gillam	0.3268	0.5410	0.6031
Kern	0.3115	0.5267	0.5690
Cruz	0.3114	0.5456	0.5966
Pavan	0.2843	0.5000	0.6055
Caurcel Diaz	0.2840	0.5000	0.5679
H. Farias	0.2816	0.5671	0.5061
Jankowska	0.2814	0.5381	0.4738
Flekova	0.2785	0.5343	0.5287
Weren	0.2564	0.5044	0.5099
Sapkota	0.2471	0.4781	0.5415
De-Arteaga	0.2450	0.4998	0.4885
Moreau	0.2395	0.4941	0.4824
baseline	0.1650	0.5000	0.3333
Gopal Patra	0.1574	0.5683	0.2895
Cagnina	0.0741	0.5040	0.1234

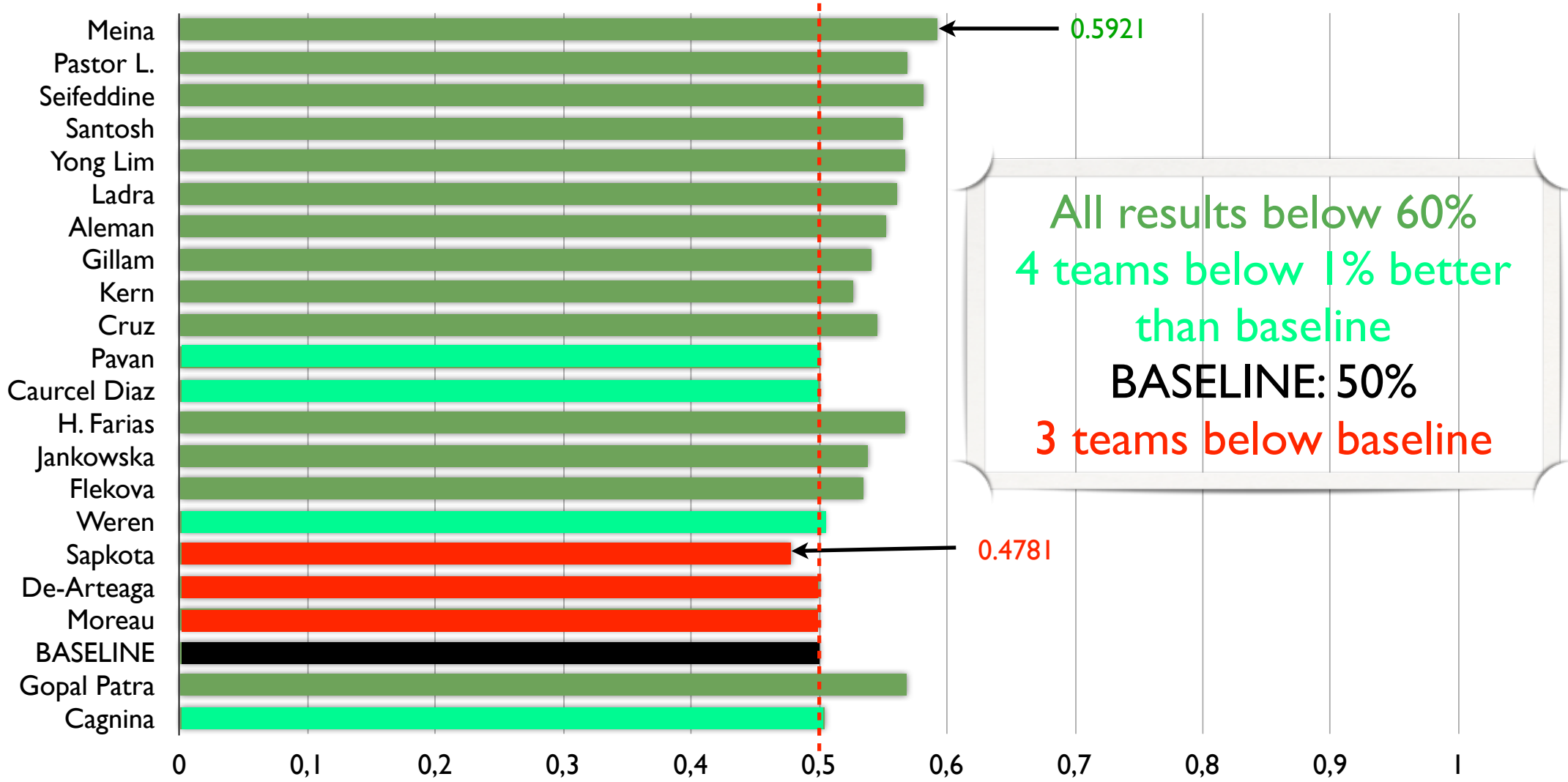
Spanish			
Team	Total	Gender	Age
Santosh	0.4208	0.6473	0.6430
Pastor L.	0.4158	0.6299	0.6558
Cruz	0.3897	0.6165	0.6219
Flekova	0.3683	0.6103	0.5966
Ladra	0.3523	0.6138	0.5727
De-Arteaga	0.3145	0.5627	0.5429
Kern	0.3134	0.5706	0.5375
Yong Lim	0.3120	0.5468	0.5705
Sapkota	0.2934	0.5116	0.5651
Pavan	0.2824	0.5000	0.5643
Jankowska	0.2592	0.5846	0.4276
Meina	0.2549	0.5287	0.4930
Gillam	0.2543	0.4784	0.5377
Moreau	0.2539	0.4967	0.5049
Weren	0.2463	0.5362	0.4615
Cagnina	0.2339	0.5516	0.4148
Caurcel Diaz	0.2000	0.5000	0.4000
H. Farias	0.1757	0.4982	0.3554
baseline	0.1650	0.5000	0.3333
Aleman	0.1638	0.5526	0.2915
Seifeddine	0.0287	0.5455	0.0512
Gopal Patra	–	–	–

- 21 teams for English, 20 teams for Spanish
- Values similar to Early Birds
  - Gender close to baseline
  - Joint identification more difficult

Results for English



Results for English



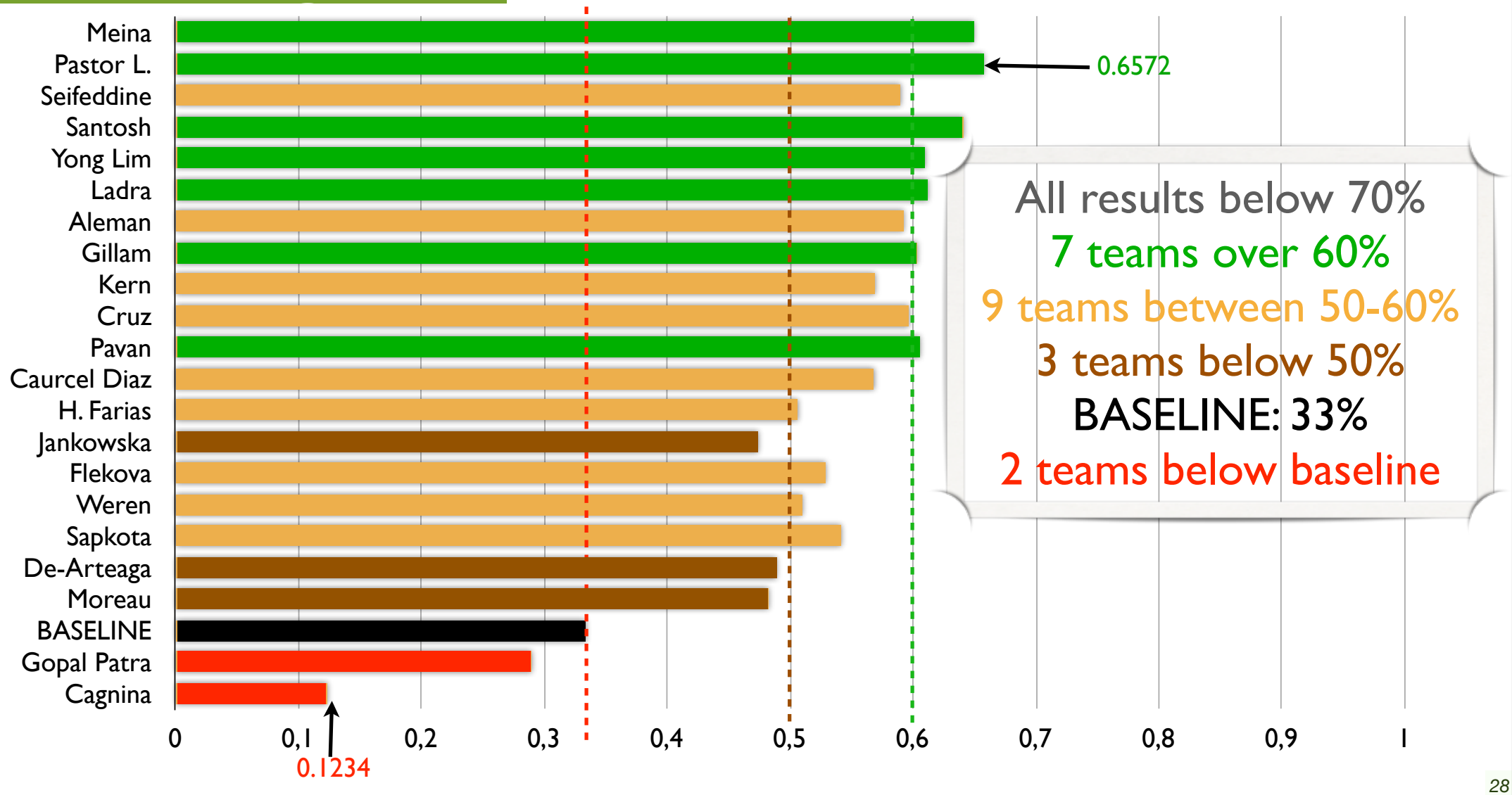
All results below 60%

4 teams below 1% better than baseline

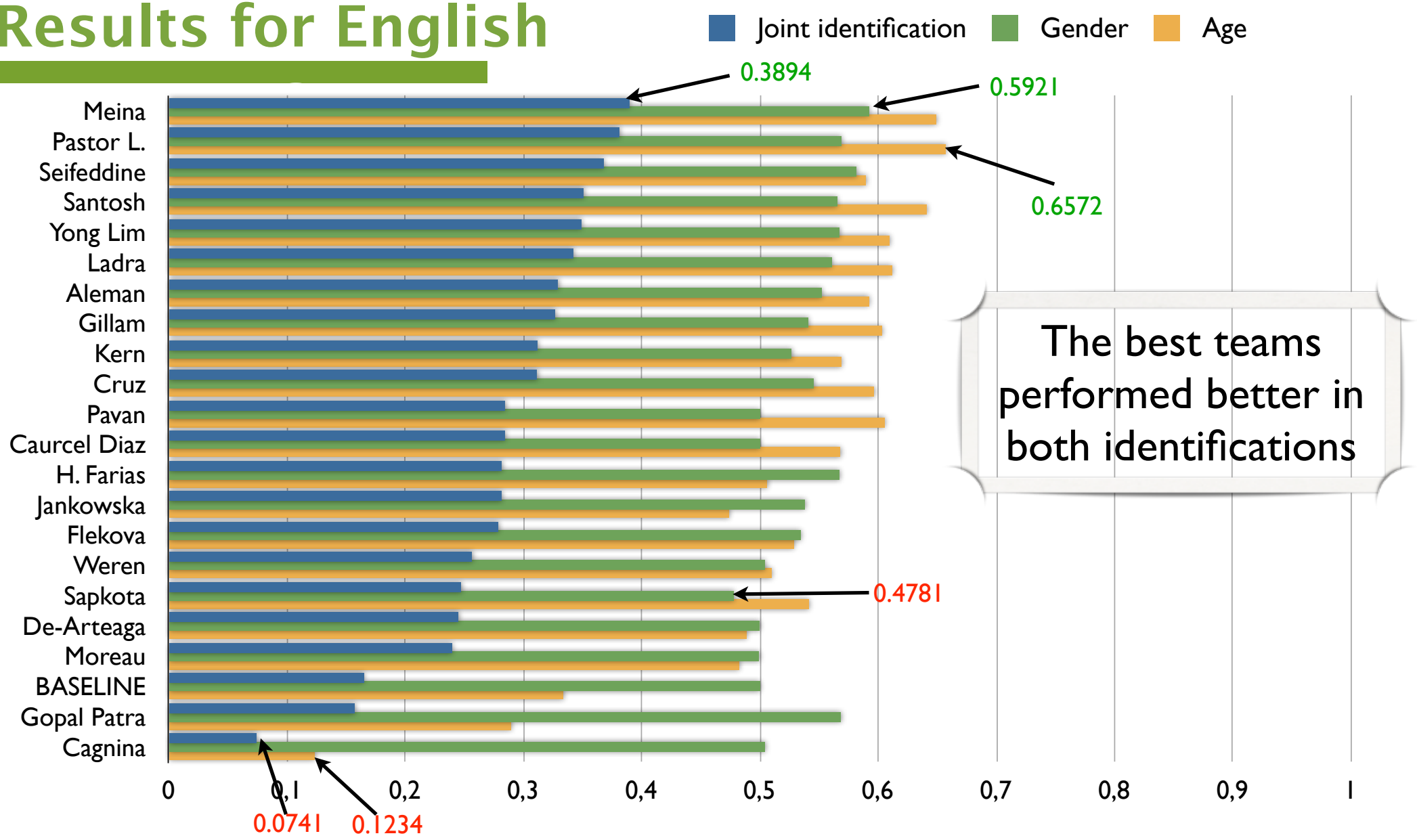
BASELINE: 50%

3 teams below baseline

Results for English

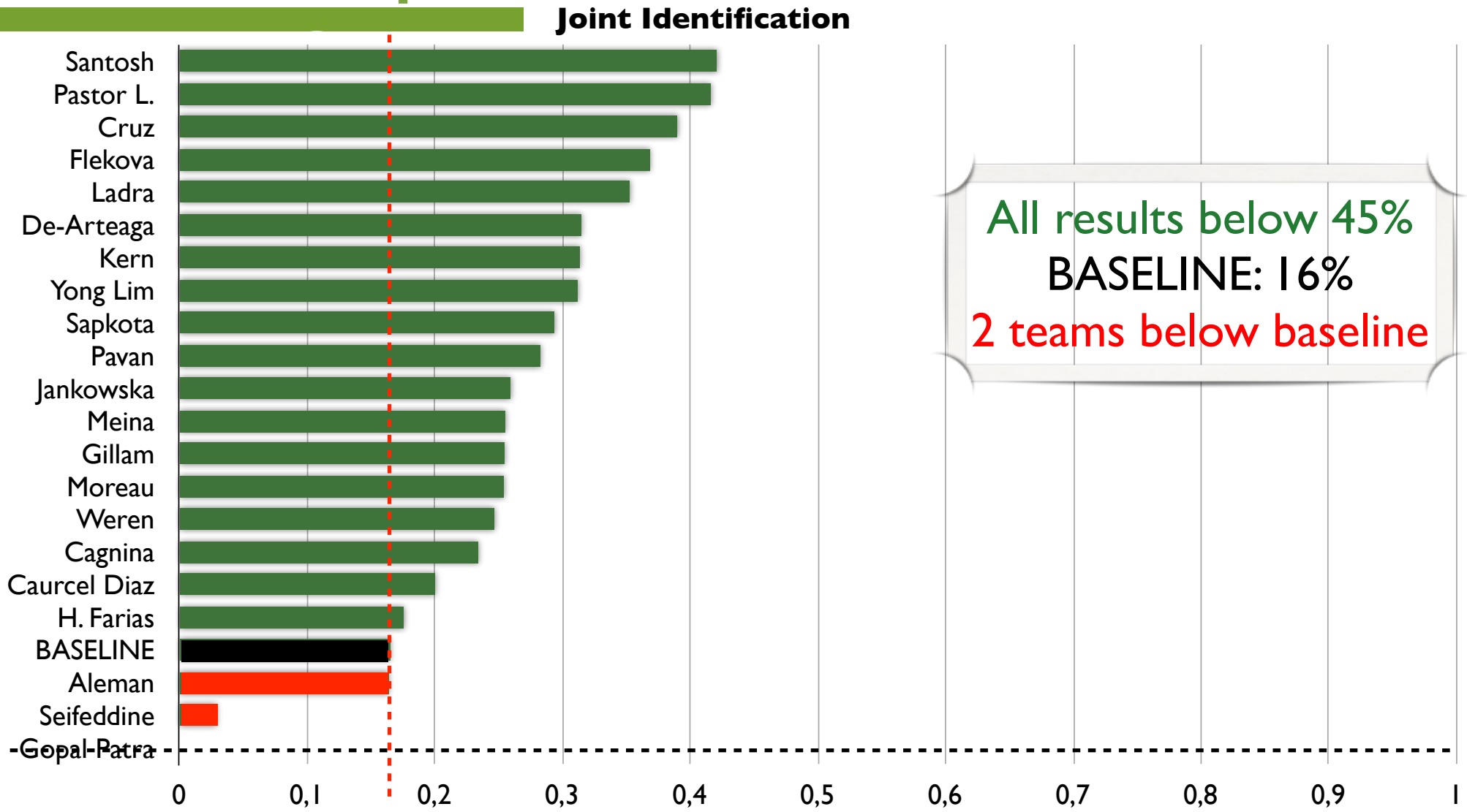


Results for English



The best teams performed better in both identifications

Results for Spanish

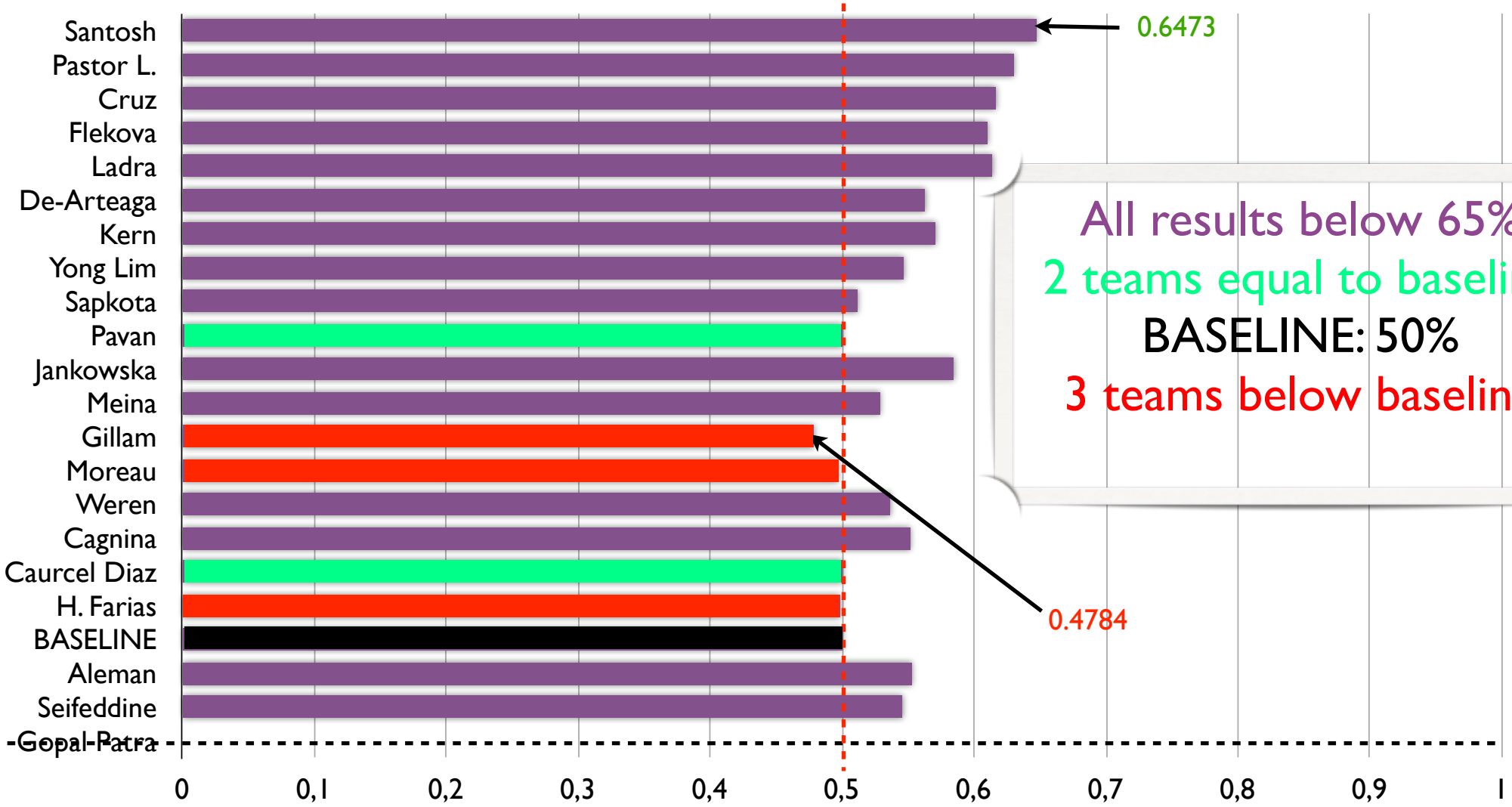


All results below 45%

BASELINE: 16%

2 teams below baseline

Results for Spanish



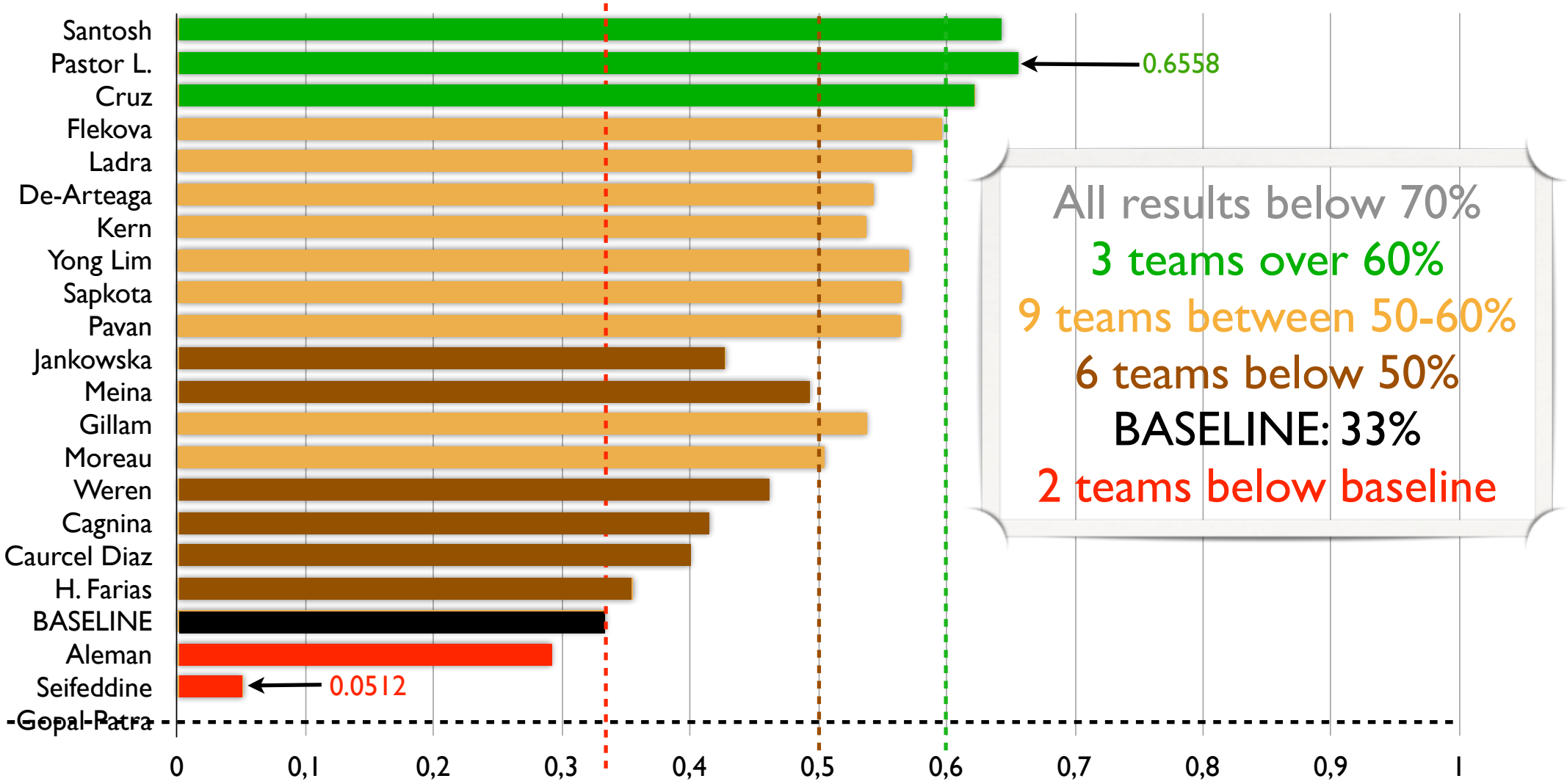
All results below 65%

2 teams equal to baseline

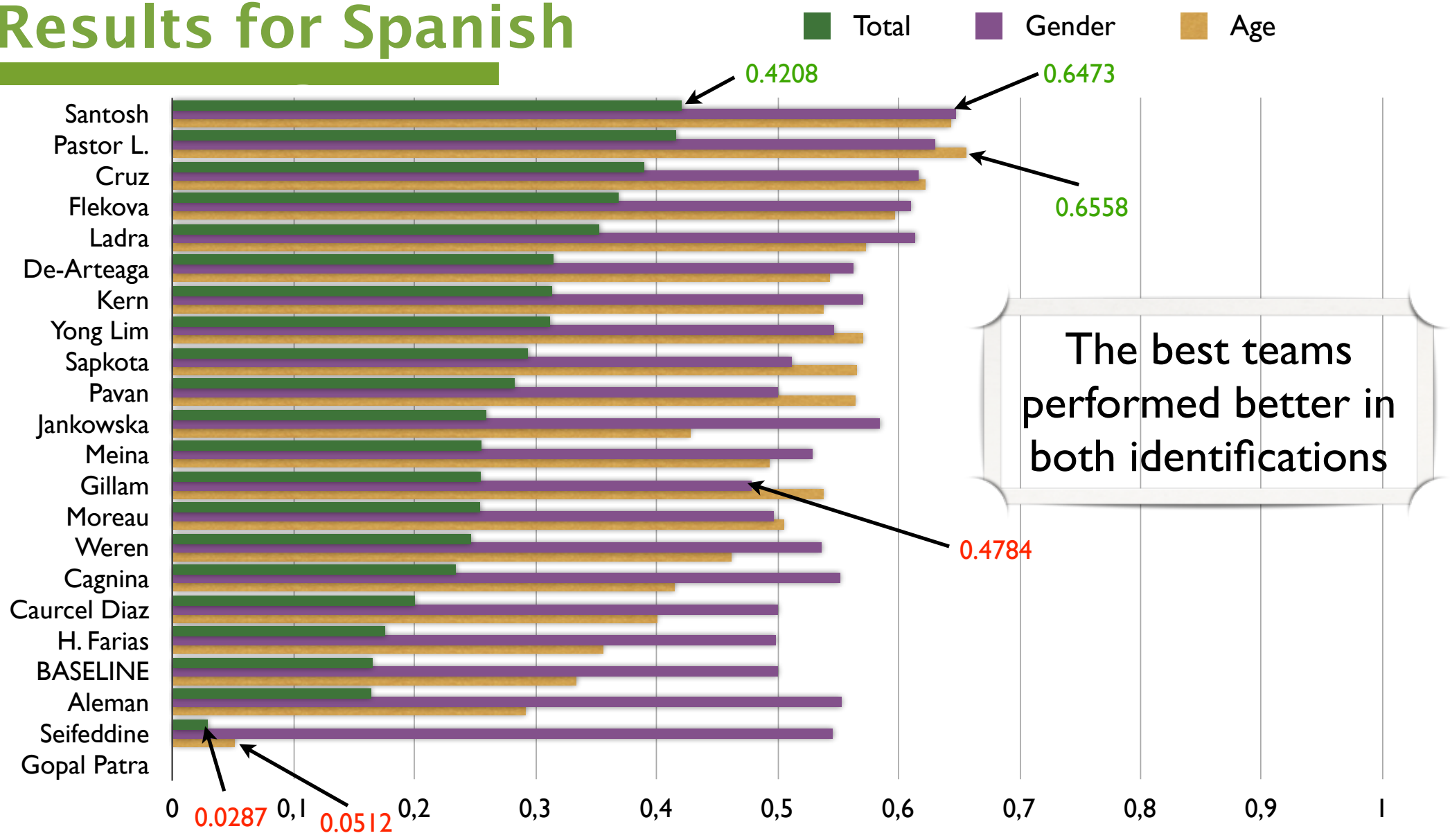
BASELINE: 50%

3 teams below baseline

Results for Spanish

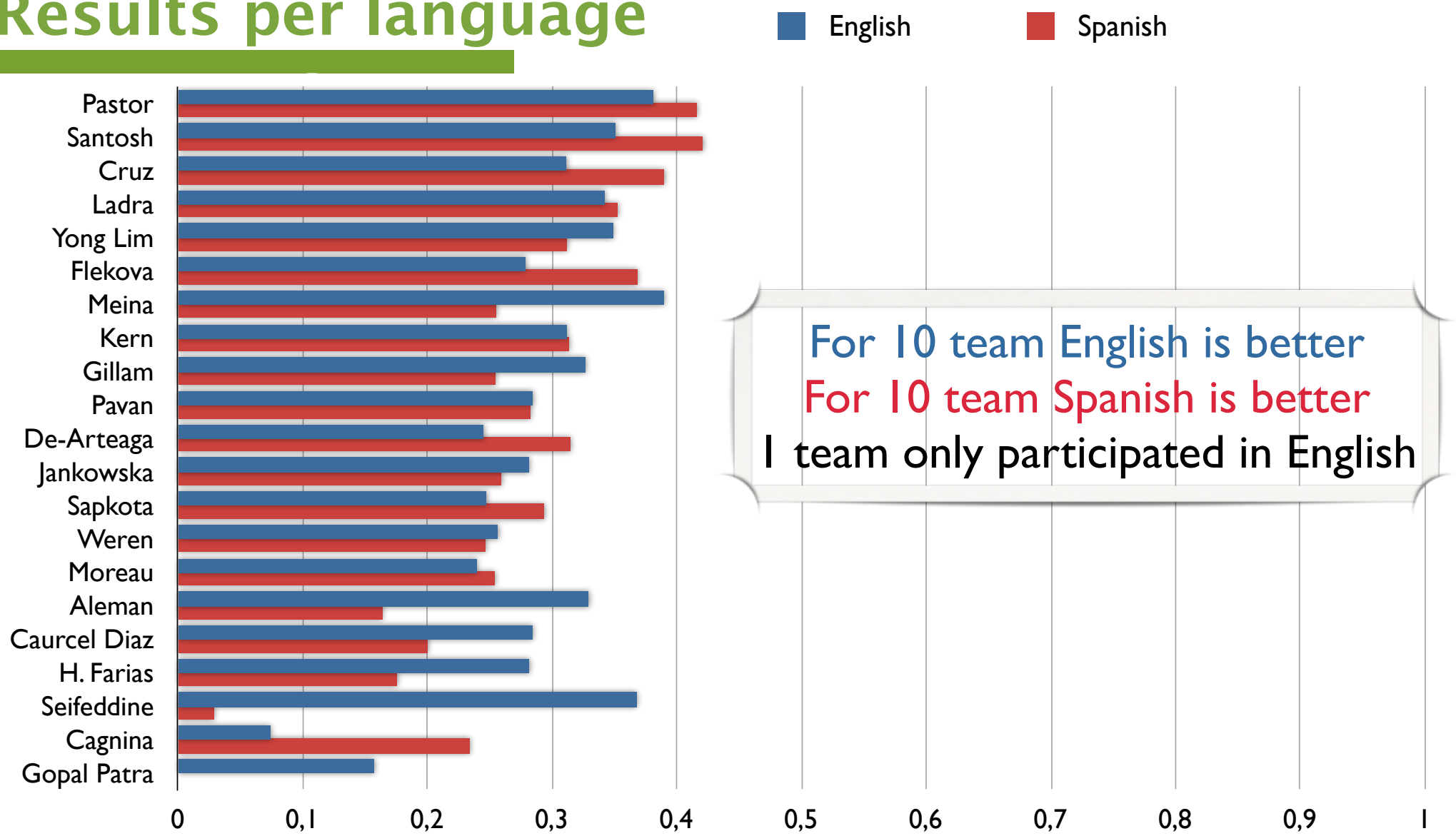


Results for Spanish



The best teams performed better in both identifications

Results per language



For 10 team English is better

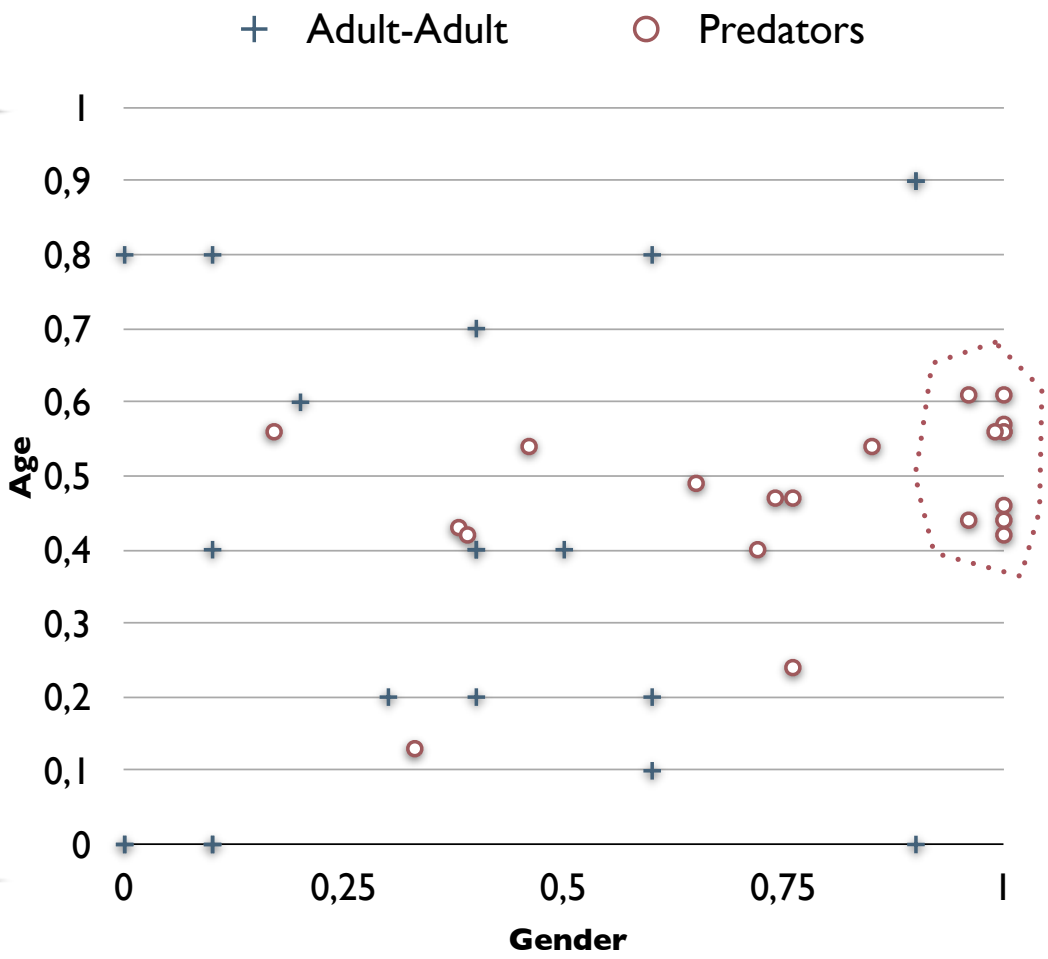
For 10 team Spanish is better

1 team only participated in English

Sexual conversations

Table 4. Number (and accuracy) of adult-adult sexual conversations (left) and predators (right) correctly identified.

Team	Adult-Adult			Predators		
	Total	Gender	Age	Total	Gender	Age
Aleman	1 (0.1)	3 (0.3)	2 (0.2)	26 (0.36)	53 (0.74)	34 (0.47)
Cagnina	4 (0.4)	4 (0.4)	7 (0.7)	8 (0.11)	24 (0.33)	9 (0.13)
Caurcel Diaz	0 (0.0)	0 (0.0)	0 (0.0)	40 (0.56)	72 (1.00)	40 (0.56)
Cruz	0 (0.0)	0 (0.0)	8 (0.8)	41 (0.57)	69 (0.96)	44 (0.61)
De Arteaga	1 (0.1)	6 (0.6)	2 (0.2)	14 (0.19)	27 (0.38)	31 (0.43)
Flekova	4 (0.4)	4 (0.4)	4 (0.4)	34 (0.47)	61 (0.85)	39 (0.54)
Gillam	0 (0.0)	1 (0.1)	4 (0.4)	30 (0.42)	72 (1.00)	30 (0.42)
Gopal Patra	1 (0.1)	5 (0.5)	4 (0.4)	12 (0.17)	55 (0.76)	17 (0.24)
H. Farias	1 (0.1)	4 (0.4)	2 (0.2)	26 (0.36)	55 (0.76)	34 (0.47)
Jankowska	0 (0.0)	1 (0.1)	0 (0.0)	44 (0.61)	72 (1.00)	44 (0.61)
Kern	9 (0.9)	9 (0.9)	9 (0.9)	25 (0.35)	47 (0.65)	35 (0.49)
Ladra	9 (0.9)	9 (0.9)	9 (0.9)	33 (0.46)	72 (1.00)	33 (0.46)
Meina	6 (0.6)	6 (0.6)	8 (0.8)	41 (0.57)	72 (1.00)	41 (0.57)
Moreau	2 (0.2)	4 (0.4)	4 (0.4)	19 (0.26)	33 (0.46)	39 (0.54)
Pastor L.	0 (0.0)	1 (0.1)	8 (0.8)	32 (0.44)	72 (1.00)	32 (0.44)
Pavan	0 (0.0)	0(0.0)	0 (0.0)	50 (0.56)	72 (1.00)	40 (0.56)
Santosh	9 (0.9)	9 (0.9)	9 (0.9)	29 (0.40)	69 (0.96)	32 (0.44)
Sapkota	0 (0.0)	9 (0.9)	0 (0.0)	9 (0.13)	12 (0.17)	40 (0.56)
Seifeddine	2 (0.2)	2 (0.2)	6 (0.6)	20 (0.28)	52 (0.72)	29 (0.40)
Weren	0 (0.0)	1 (0.1)	0 (0.0)	39 (0.54)	71 (0.99)	40 (0.56)
Yong Lim	1 (0.1)	6 (0.6)	1 (0.1)	17 (0.24)	28 (0.39)	30 (0.42)



- Age identification similar to previous values
- Gender identification performed better in case of predators

Runtime

- Big Data problem?

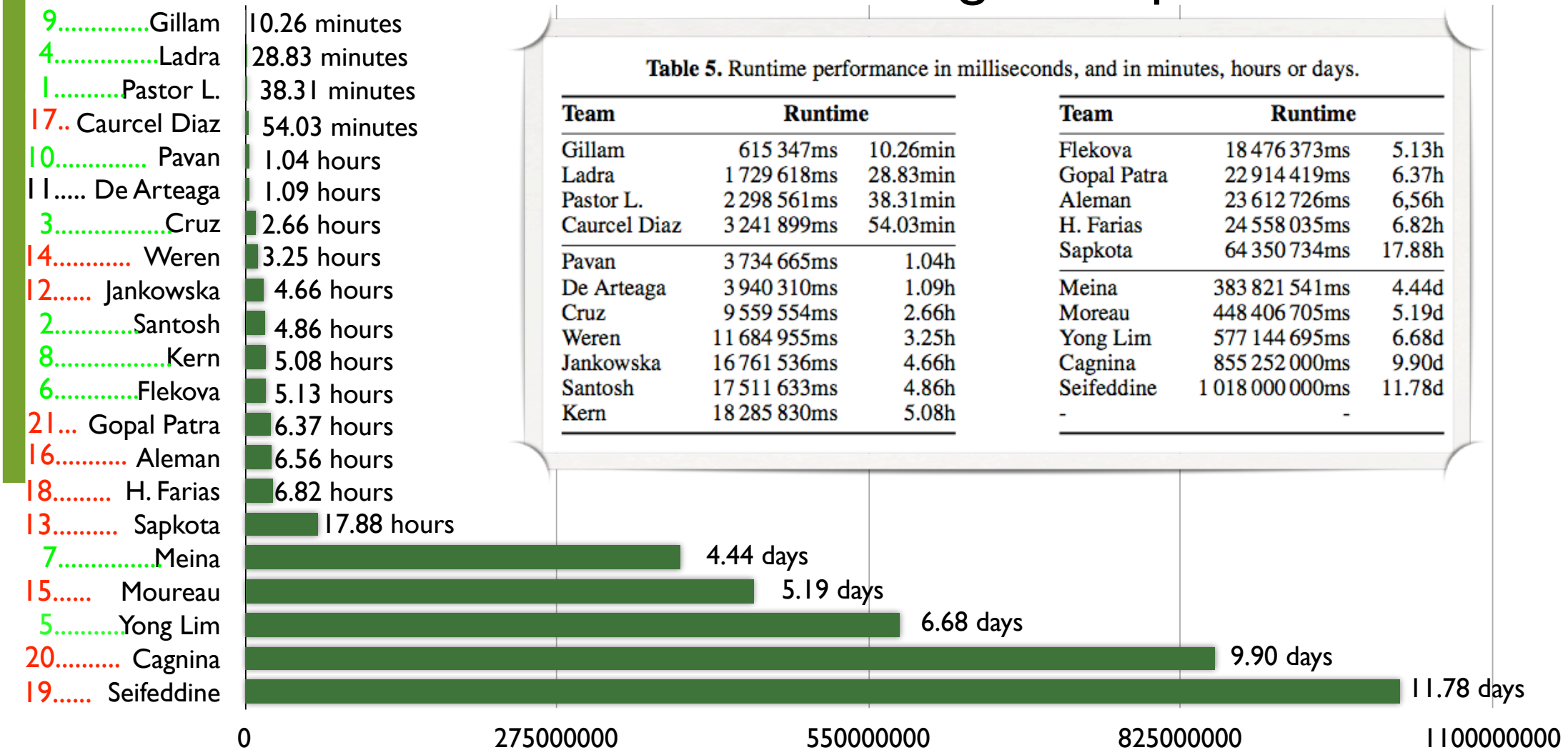


Table 5. Runtime performance in milliseconds, and in minutes, hours or days.

Team	Runtime	
Gillam	615 347ms	10.26min
Ladra	1 729 618ms	28.83min
Pastor L.	2 298 561ms	38.31min
Caurcel Diaz	3 241 899ms	54.03min
Pavan	3 734 665ms	1.04h
De Arteaga	3 940 310ms	1.09h
Cruz	9 559 554ms	2.66h
Weren	11 684 955ms	3.25h
Jankowska	16 761 536ms	4.66h
Santosh	17 511 633ms	4.86h
Kern	18 285 830ms	5.08h

Team	Runtime	
Flekova	18 476 373ms	5.13h
Gopal Patra	22 914 419ms	6.37h
Aleman	23 612 726ms	6,56h
H. Farias	24 558 035ms	6.82h
Sapkota	64 350 734ms	17.88h
Meina	383 821 541ms	4.44d
Moreau	448 406 705ms	5.19d
Yong Lim	577 144 695ms	6.68d
Cagnina	855 252 000ms	9.90d
Seifeddine	1 018 000 000ms	11.78d
-	-	

Conclusions

- ***Very difficult task, mainly for gender identification***
- ***Difficult to identify together age and gender***

For predators...

- Robust identifying age
- Better identifying gender

- Expensive in Time consuming -> Big Data problem?

Also...

- We received many different and enriching approaches
- We were one of the task with the higher number of participants at CLEF (21)
- Interest from many teams (66 registered) but the task was new and many (5) did not make it (potentially more participation next year!)

Francisco Rangel
Autoritas Consulting /
Universitat Politècnica de València

Paolo Rosso
Universitat Politècnica de València

Moshe Koppel
Bar-Ilan University

Efstathios Stamatatos
University of the Aegean

Giacomo Inches
University of Lugano

On behalf of the AP task organisers:
Thank you very much for participating!
We hope to see you again next year!

autoritas
nuevas ideas, nuevas soluciones

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Tradition of Excellence
Bar-Ilan University
אוניברסיטת בר-אילן

UNIVERSITY OF THE AEGEAN

Università della Svizzera italiana