Author Profiling in Social Media

Paolo Rosso

NLE Lab, Universitat Politècnica de València

NLEL
Natural Language Engineering Lab

RepLab @ CLEF-2013
Valencia, 24th September 2013

Francisco Rangel
Autoritas / 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	Marketing	Security
Language as evidence	***Profiling costumers***	***Profiling delinquents***

Sexual predators

- Diagnostic and Statistical Manual of Mental Disorders: *A pedophile is an individual who fantasizes about, is sexually aroused by, or experiences sexual urges toward prepubescent children (generally <13 years) for a period of at least 6 months*
- 88% of child sexual molesters are pedophiles
- 67% of sexual assault victims are underaged
- 19% of children have been sexually approached over the Internet

Perverted Justice

- **Perverted Justice Foundation** investigates and publishes cases of online sexual predators
- **Adult volunteers enter chat rooms as children.** If they are approached they pass information to the police
- The chat data is available at http://perverted-justice.com
- **Myspace** (10,786) known sex offenders since 2007); **Facebook**: 2,800 since 2008.

Sexual Predators in Twitter

Mirror NEWS

Real news, real entert

FRONT PAGE | NEWS | SPORT | 3AM | TV | LIFESTYLE | MONEY

News · UK News · London 2012 Olympics

By Dominic Herbert | 12 Aug 2012 00:27

Twitter paedos exposed: Vile perverts using social networking site to find victims and trade intelligence

Within two minutes of searching we found 20 paedophiles wanting to abuse young children - and 200 in two hours

- http://www.mirror.co.uk/news/uk-news/paedophiles-using-twitter-to-find-victims-1253833

Analysis of Fake Profiles in Social Networks

- http://www.reuters.com/article/2012/07/12/us-usa-internet-predators-idUSBRE86B05G20120712

Sexual Predators Identification

- Competition at PAN-2012
- 16 participants

Giacomo Inches
University of Lugano

Fabio Crestani
University of Lugano

Outline of the Talk

1. Age & Gender Identification
2. Author Profiling at PAN@CLEF-2013
3. Personality Traits & Native Language Identification

1. Age & Gender Identification

- Related Work

Moshe Koppel
Bar-Ilan University

Which is Male&Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

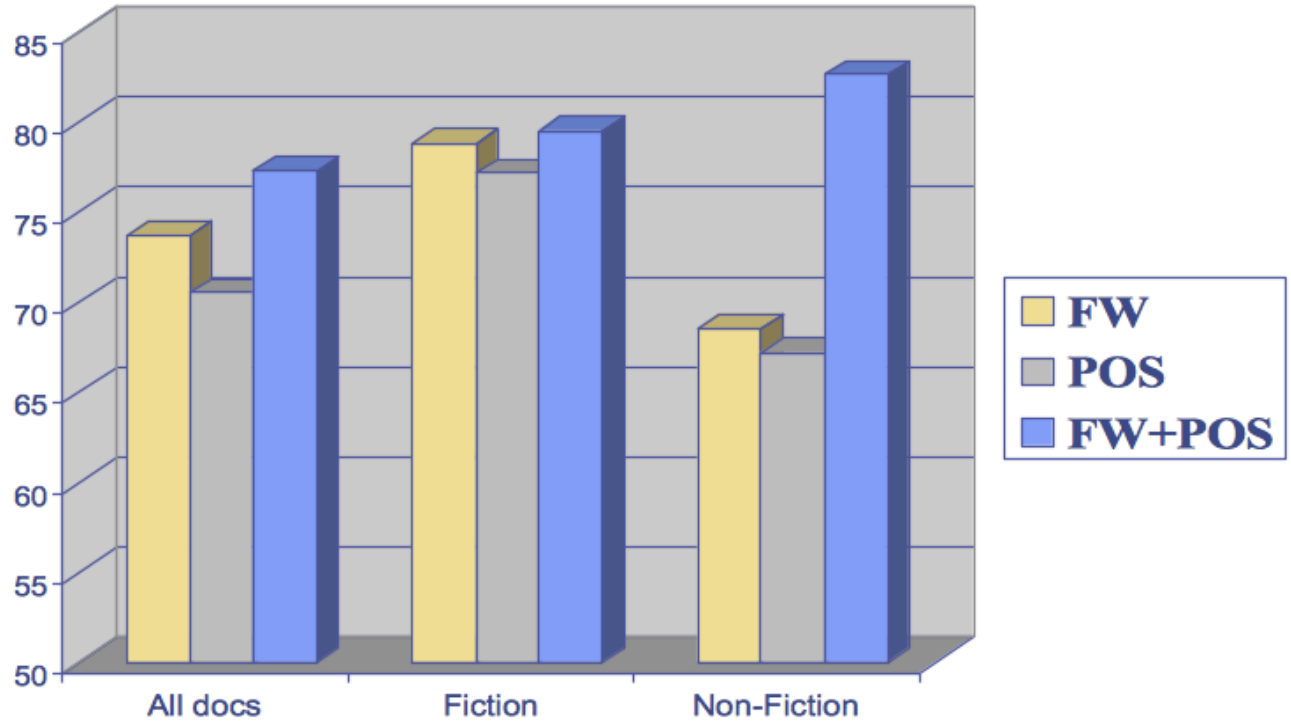
The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

British National Corpus

- 920 documents labelled for
  - author gender
  - document genre
- Used 566 controlled for genre

	Male	Fem
Fiction (prose)	132	132
Non-fiction	151	151
Arts (general)	8	8
Arts (acad.)	12	12
Belief/Thought	12	12
Biography	27	27
Commerce	5	5
Leisure	8	8
Science (gen.)	13	13
Soc. Sci. (gen.)	26	26
Soc. Sci. (acad.)	19	19
World Affairs	21	21

Results per Feature Set



- Handle fiction and non-fiction separately
- Use full feature set

POS: Part-Of-Speech FW: Function words (*and, of, the,..*)

What are the Distinguishing Features?

- Fiction
  - Male: *a, the, as*
  - Female: *she, for, with, not*
- Non-Fiction
  - Male: *that, one, of,* **preposition**, **article**
  - Female: *she, for, with, and, in,* **pronoun**

Summary: Male vs. Female Style

Males use more

- Determiners
- Adjectives
- *of* modifiers (e.g. *pot of gold*)

Informational features

- Females use more
- Pronouns
- *for* and *with*
- Negation
- Present tense

Involvedness features

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Which is Male/Female?

My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding of what some of these so-called apposition markers indicate. It will be argued that the decision to put something in other words is essentially a decision about style, a point which is, perhaps, anticipated by Burton-Roberts when he describes loose apposition as a rhetorical device. However, he does not justify this suggestion by giving the criteria for classifying a mode of expression as a rhetorical device. Nor does he specify what kind of effects might be achieved by a reformulation or explain how it achieves those effects. In this paper I follow Sperber and Wilson's (1986) suggestion that rhetorical devices like metaphor, irony and repetition are particular means of achieving relevance. As I have suggested, the corrections that are made in unplanned discourse are also made in the pursuit of optimal relevance. However, these are made because the speaker recognises that the original formulation did not achieve optimal relevance .

The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. It details the design and results of a workshop activity on narrative carried out with undergraduates in a university department of English. The methods proposed are intended to enable students to obtain insights into aspects of cohesion and narrative structure: insights, it is suggested, which are not as readily obtainable through more traditional techniques of stylistic analysis. The text chosen for analysis is a short story by Ernest Hemingway comprising only 11 sentences. A jumbled version of this story is presented to students who are asked to assemble a cohesive and well formed version of the story. Their re-constructions are then compared with the original Hemingway version.

Blog Corpus

- Less-formal text
  - 85,000 blogs
  - blogger-provided profiles (gender, age, occupation, astrological sign)
  - harvested August 2004
  - all non-text ignored (formatting, quoting)

Example 1

Thirties Twenties Teen

Female Male

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotton, and I wanted to cry, but...it's .ok

Example 1

Thirties Twenties **Teen**

Female Male

Yesterday we had our second jazz competition. Thank God we weren't competing. We were sooo bad. Like, I was so ashamed, I didn't even want to talk to anyone after. I felt so rotton, and I wanted to cry, but...it's .ok

Example 2

Thirties Twenties Teen

Female Male

My gracious boss had agreed to let me have one week off of "work." He did finally give me my report back after eight freakin' days! Now I only have the rest of this week and then one full week after my vacation .to finish this damned thing

Example 2

Thirties Twenties Teen

Female Male

My gracious boss had agreed to let me have one week off of "work." He did finally give me my report back after eight freakin' days! Now I only have the rest of this week and then one full week after my vacation .to finish this damned thing

Blog Corpus

	Gender		
Age	Female	Male	Total
~~unknown~~	~~12287~~	~~12259~~	~~24546~~
13-17	~~6949~~	***4120***	8240
~~18-22~~	~~7393~~	~~7690~~	~~15083~~
23-27	***4043***	~~6062~~	8086
~~28-32~~	~~1686~~	~~3057~~	~~4743~~
33-37	***860***	~~1827~~	1720
38-42	***374***	~~819~~	748
43-48	***263***	~~584~~	526
~~>48~~	~~314~~	~~906~~	~~1220~~
Total	**9660**	**9660**	**19320**

Final balanced corpus:

- 19,320 total blogs
  - 8240 in "10s"
  - 8086 in "20s"
  - 2994 in "30s"
- 681,288 total posts
- 141,106,859 total words

Experimental Setup

Feature sets:

- Content: words (filtered by Info Gain on train set)
- Style: parts-of-speech, function words, blog slang

Learning algorithms:

- Real-valued balanced winnow (RBW)
- Bayesian multinomial regression (BMR)

Evaluation: 10-fold cross-validation

Age Classification

	RBW	BMR
Style & Content	75.0%	77.4%
Style Words	67.7%	69.4%
Content Words	75.9%	76.2%

The Lifecycle of the Common Blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

The Lifecycle of the Common Blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

The Lifecycle of the Common Blogger...

Word	10s	20s	30s
maths	105	3	2
homework	137	18	15
bored	384	111	47
sis	74	26	10
boring	369	102	63
awesome	292	128	57
mum	125	41	23
crappy	46	28	11
mad	216	80	53
dumb	89	45	22

Word	10s	20s	30s
semester	22	44	18
apartment	18	123	55
drunk	77	88	41
beer	32	115	70
student	65	98	61
album	64	84	56
college	151	192	131
someday	35	40	28
dating	31	52	37
bar	45	153	111

Word	10s	20s	30s
marriage	27	83	141
development	16	50	82
campaign	14	38	70
tax	14	38	72
local	38	118	185
democratic	13	29	59
son	51	92	237
systems	12	36	55
provide	15	54	69
workers	10	35	46

Gender Classification

	RBW
Style & Content	80.0%
Style Words	77.0%
Content Words	73.0%

Male are from Mars / Female are from Venus

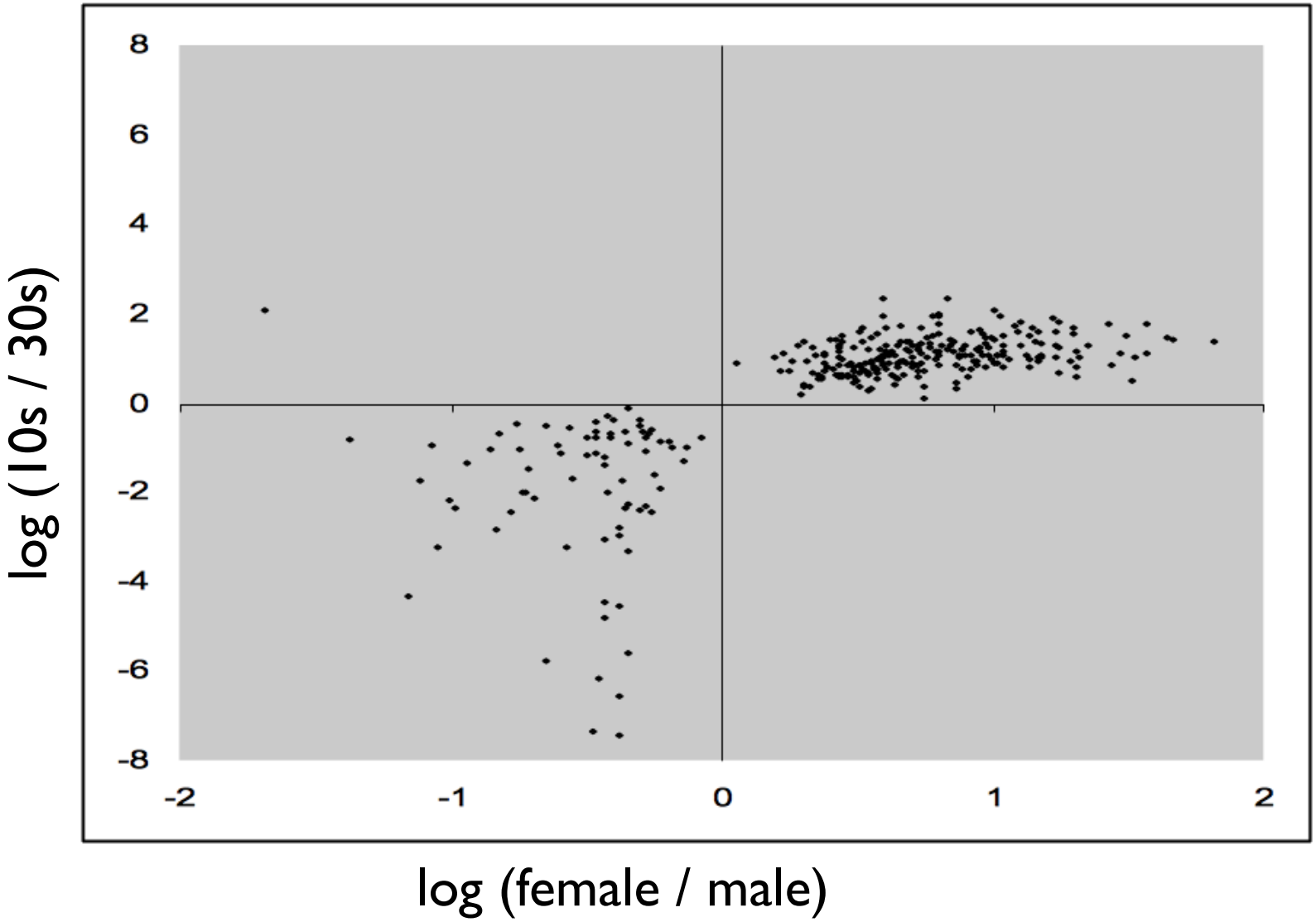
LIWC category	male	female
job	**68.1±0.6**	56.5±0.5
money	**43.6±0.4**	37.1±0.4
sports	**31.2±0.4**	20.4±0.2
tv	**21.1±0.3**	15.9±0.2
sex	32.4±0.4	**43.2±0.5**
family	27.5±0.3	**40.6±0.4**
eating	23.9±0.3	**30.4±0.3**
friends	20.5±0.2	**25.9±0.3**
sleep	18.4±0.2	**23.5±0.2**
pos-emotions	*248.2±1.9*	*265.1±2*
neg-emotions	*159.5±1.3*	*178±1.4*

LIWC: Linguistic Inquiry and Word Count [J.W. Pennebaker]

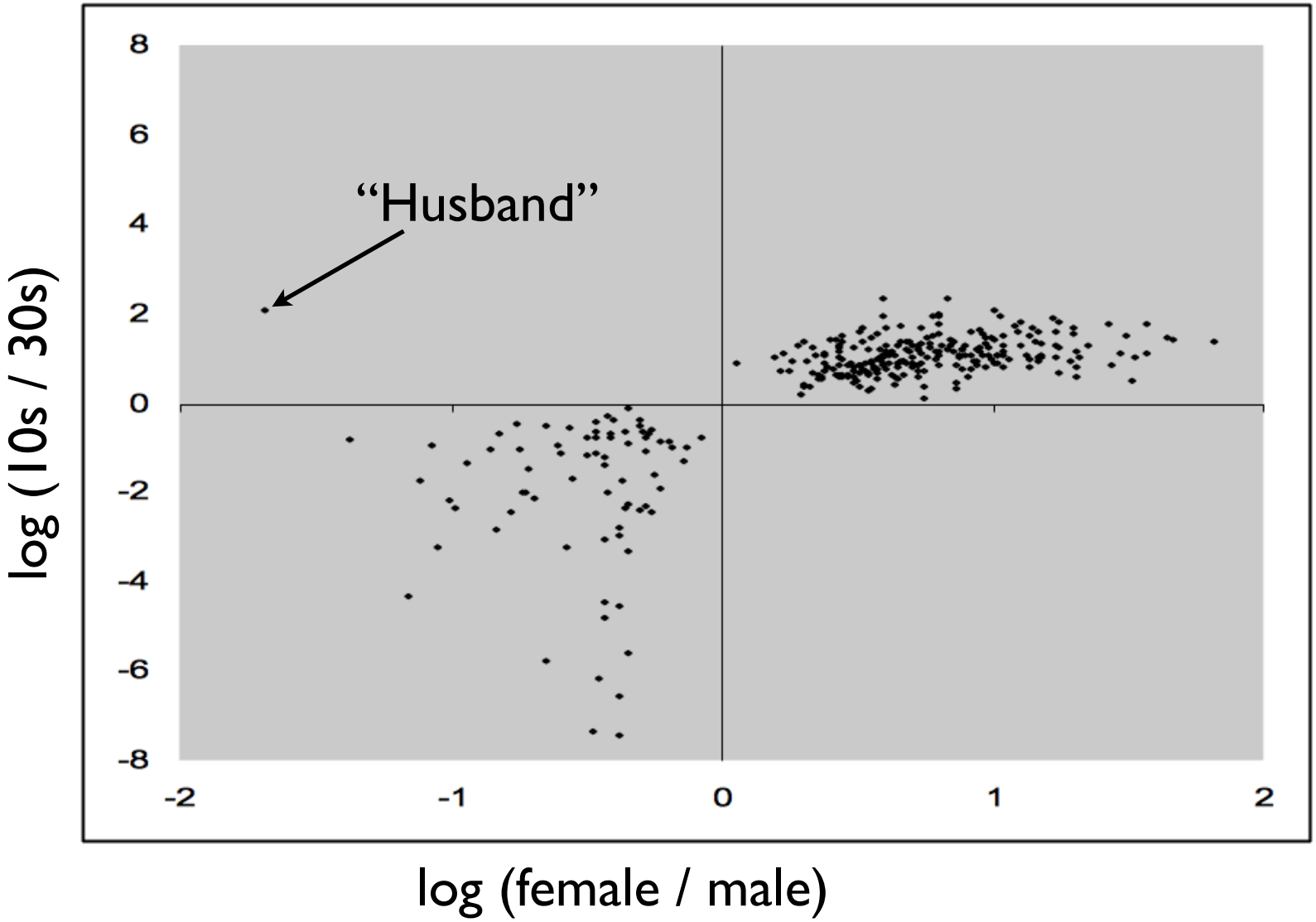
Relating Age and Gender

- Now…is there a linguistic connection between age and gender?
- Consider the most distinctive words for both Age and Gender:
  - Intersect the 1000 words with **highest Age information gain** and the 1000 words with **highest Gender information gain**
  - Total of 316 words
  - Plot log(30s/10s) vs. log(male/female)

Relating Age and Gender



Relating Age and Gender



More 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

2. Author Profiling at PAN@CLEF-2013

PAN

CLEF 2013 Valencia

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

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

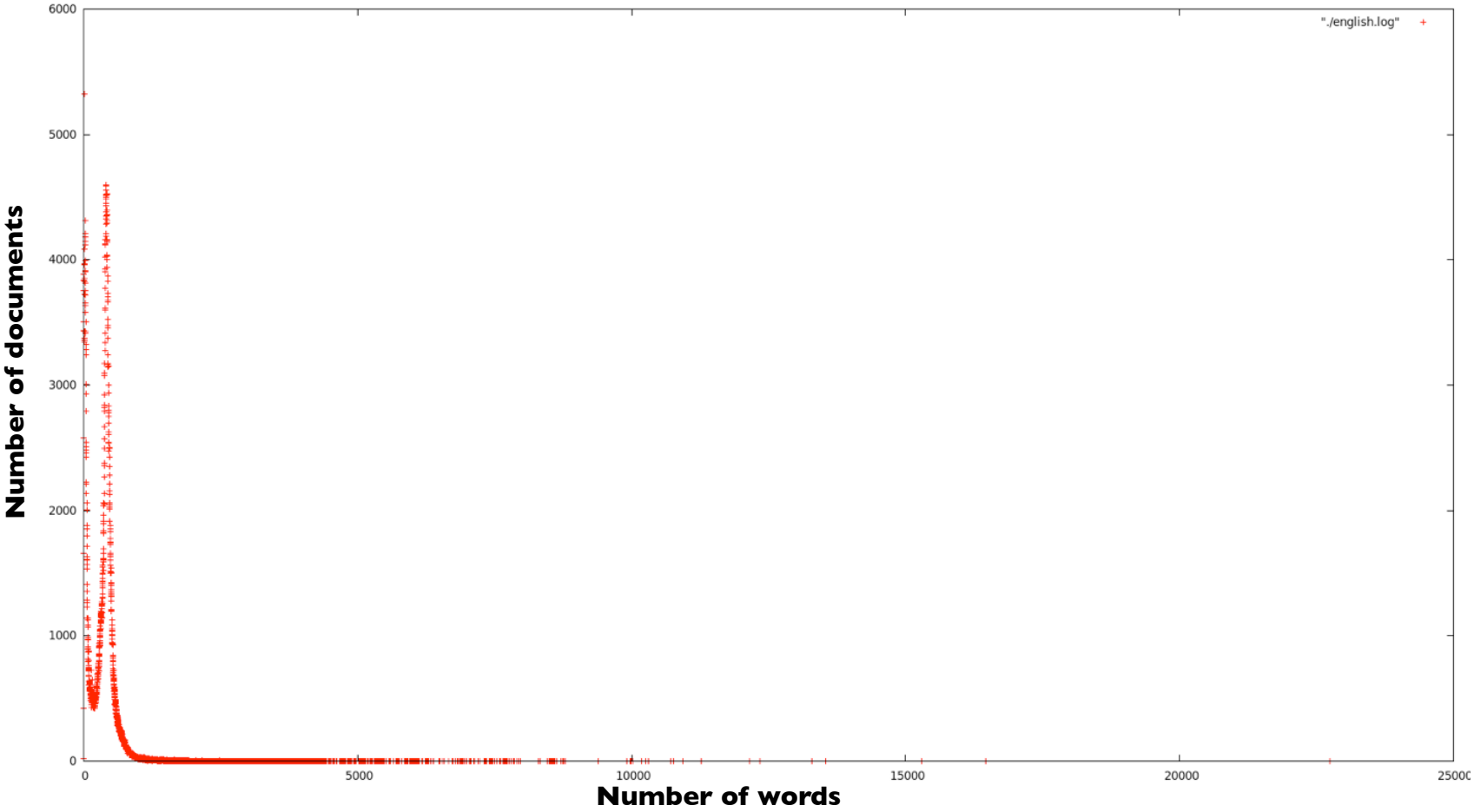
Data Collection – Social Media

- 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



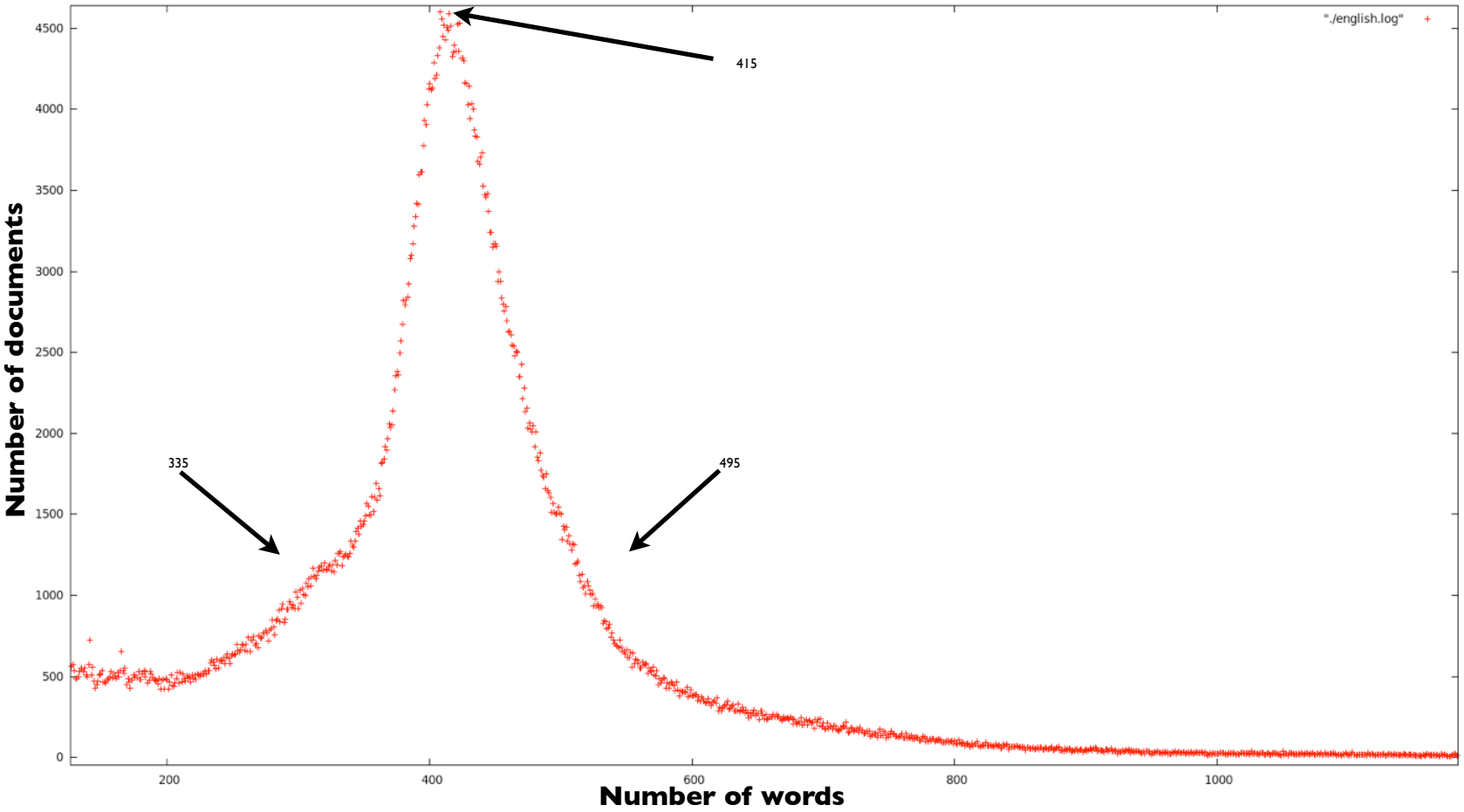



Data Collection – English Distribution



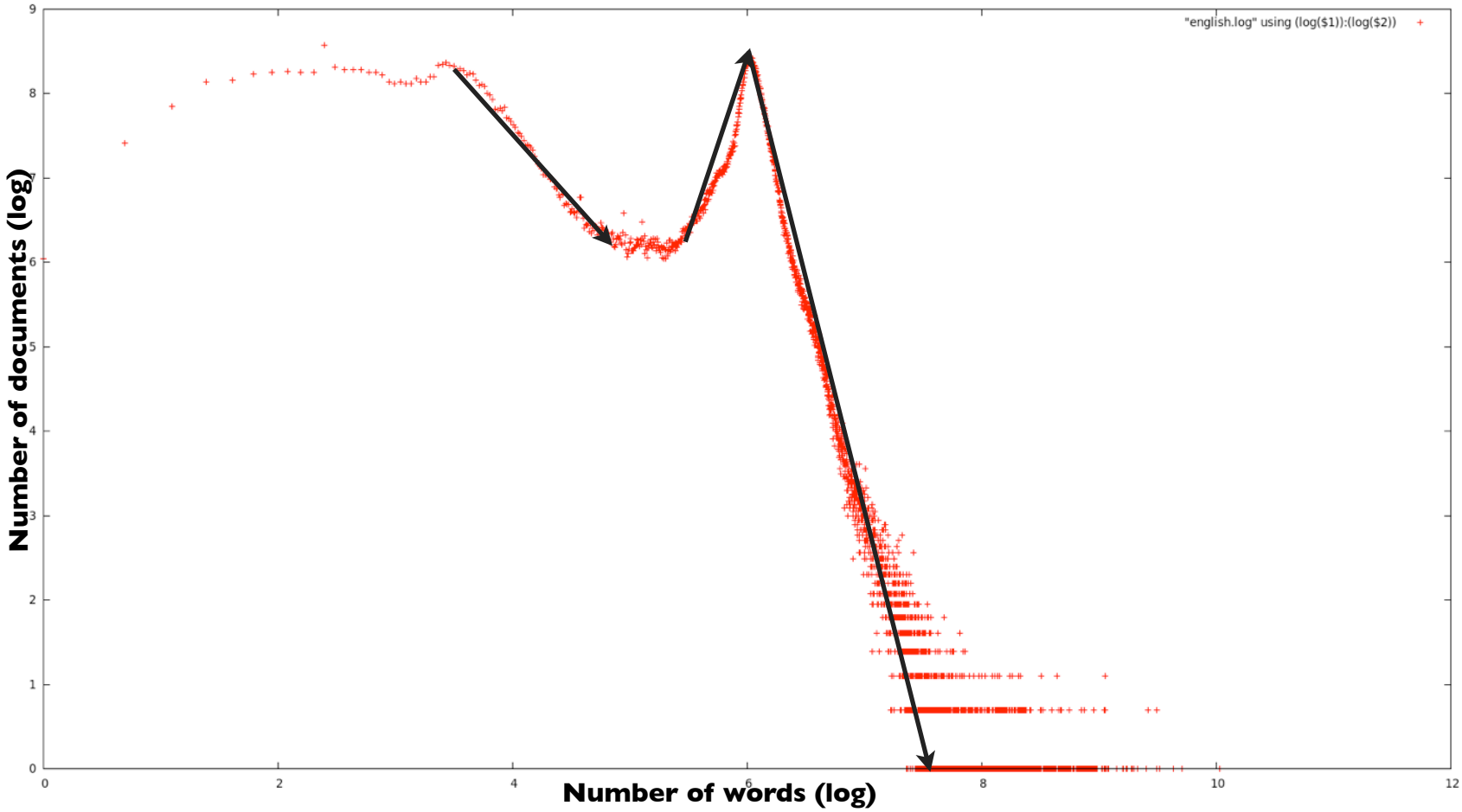
MIN	MAX	AVG	STD
0	22,736	335	208

Data Collection – English Distribution (zoomed)



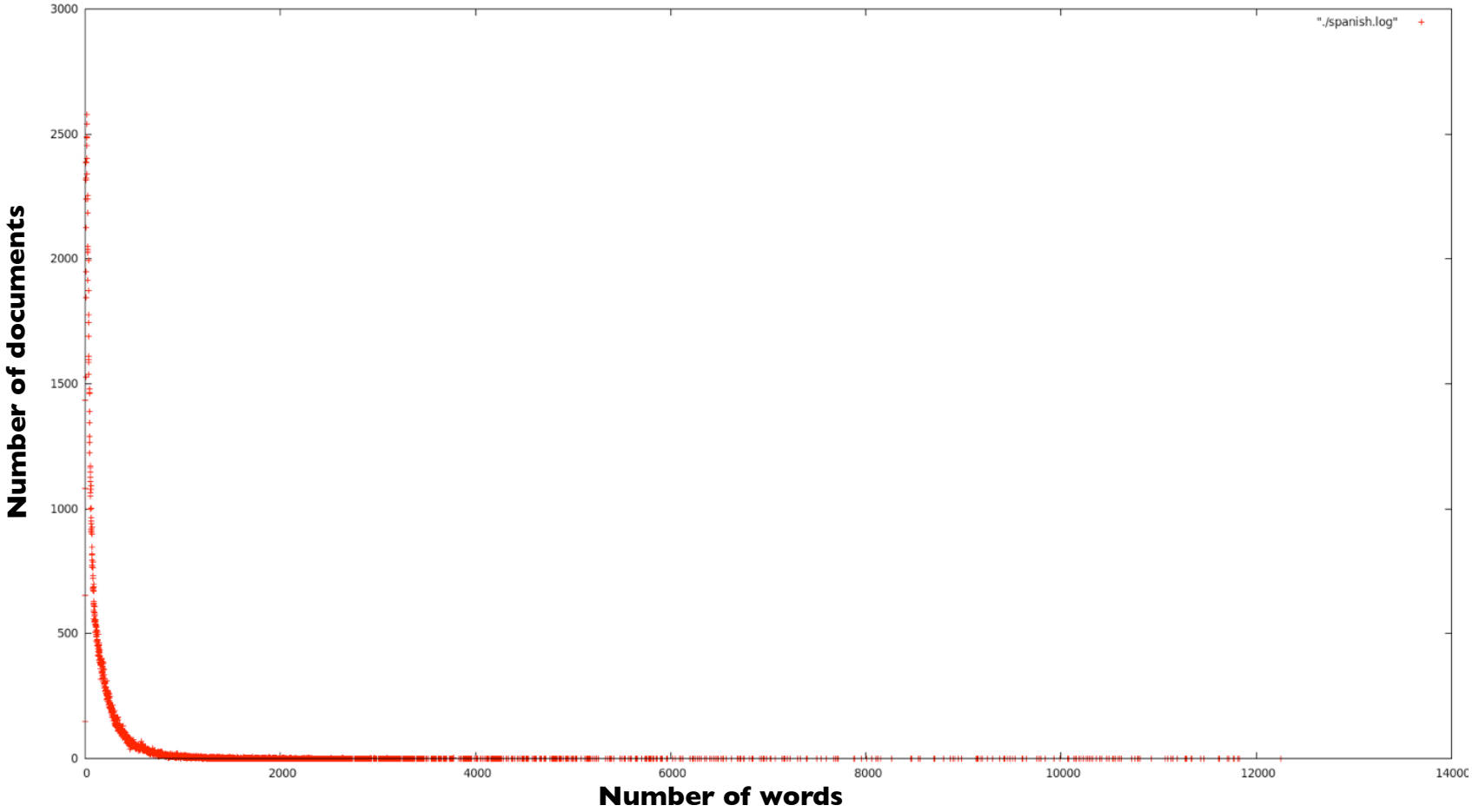
- If we zoom the distribution, we can observe a gaussian like distribution, with its maximum on the value 415.

Data Collection – English Distribution (log-log)



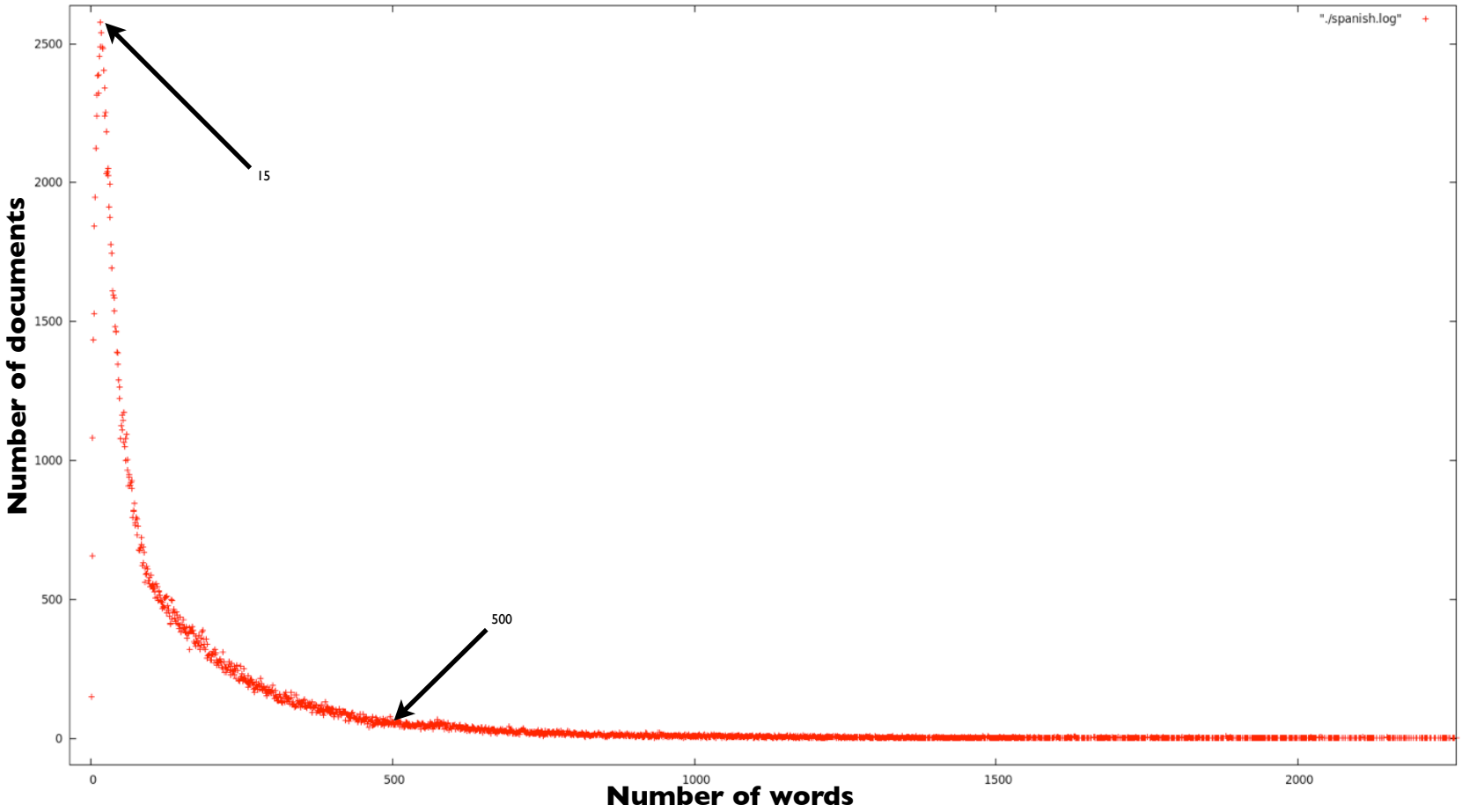
- The log-log representation shows how the distribution has a long tail component, specifically in two cases, one before the point of maximum frequency and another one after this.
- We could use this property to select minimum and maximum number of words that the posts must have.

Data Collection – Spanish Distribution

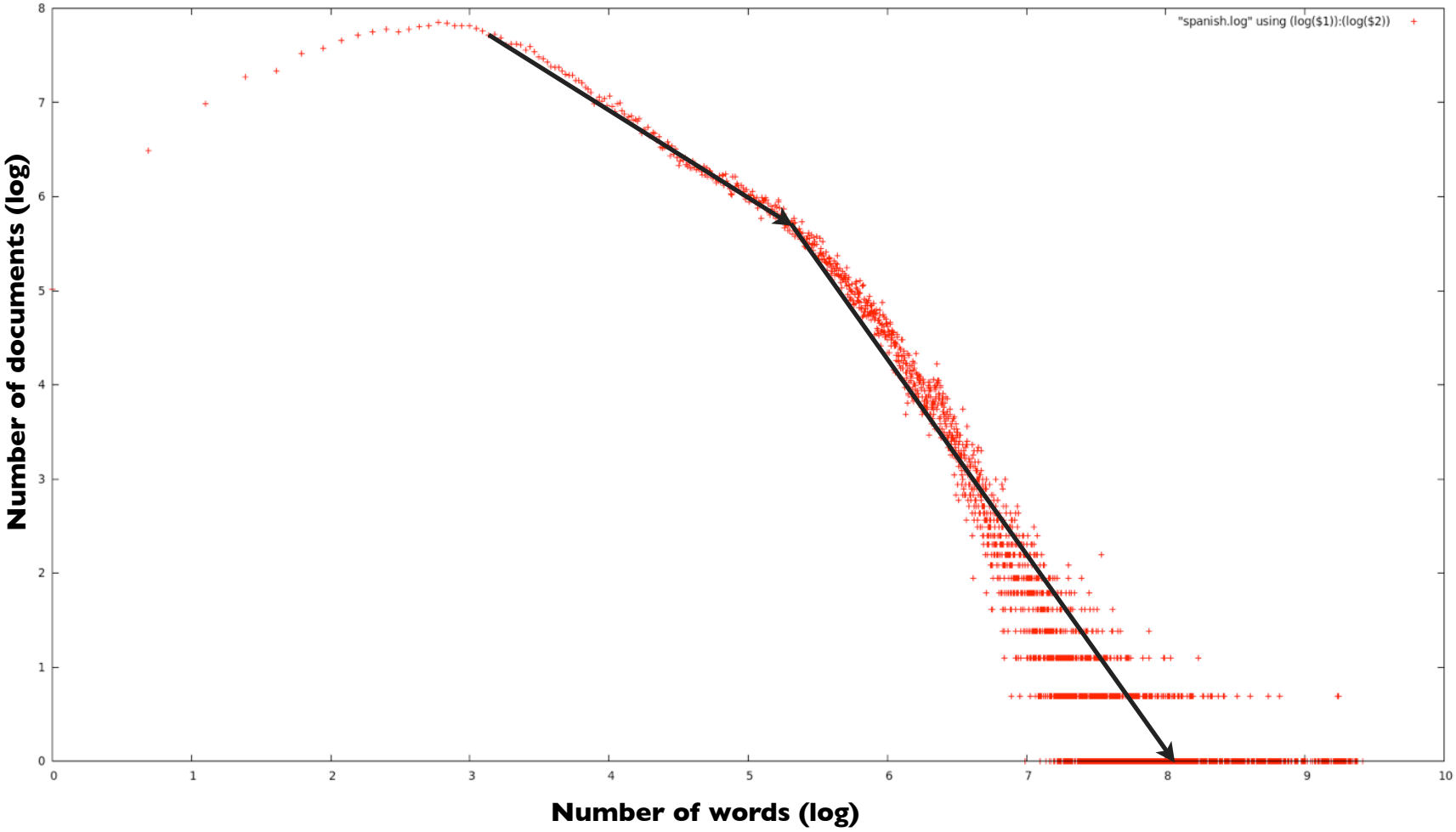


MIN	MAX	AVG	STD
0	12,246	176	832

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%)

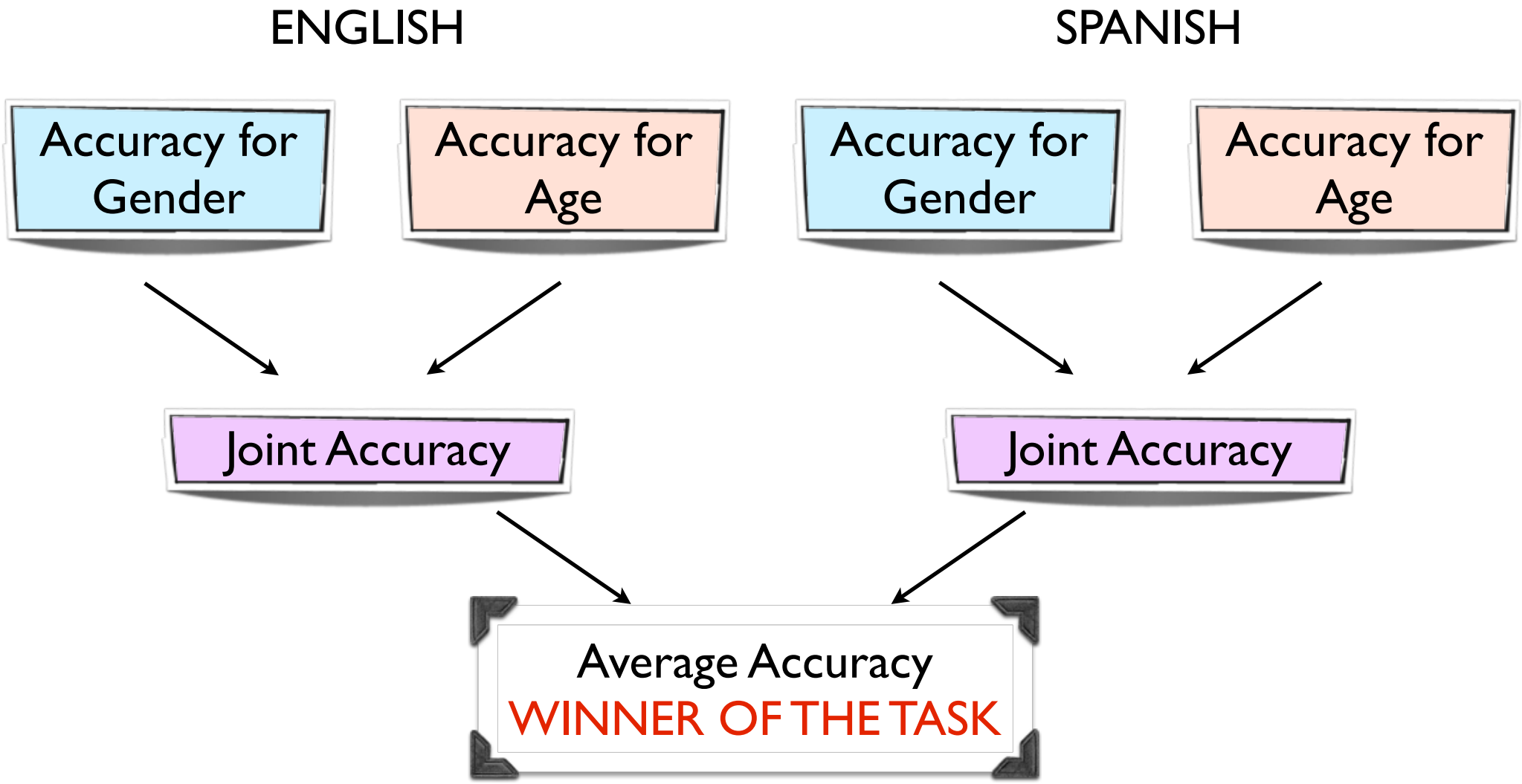
- Introduction of few special cases
  - Predators (0.0012%)
  - Adult-adult sexual conversations

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



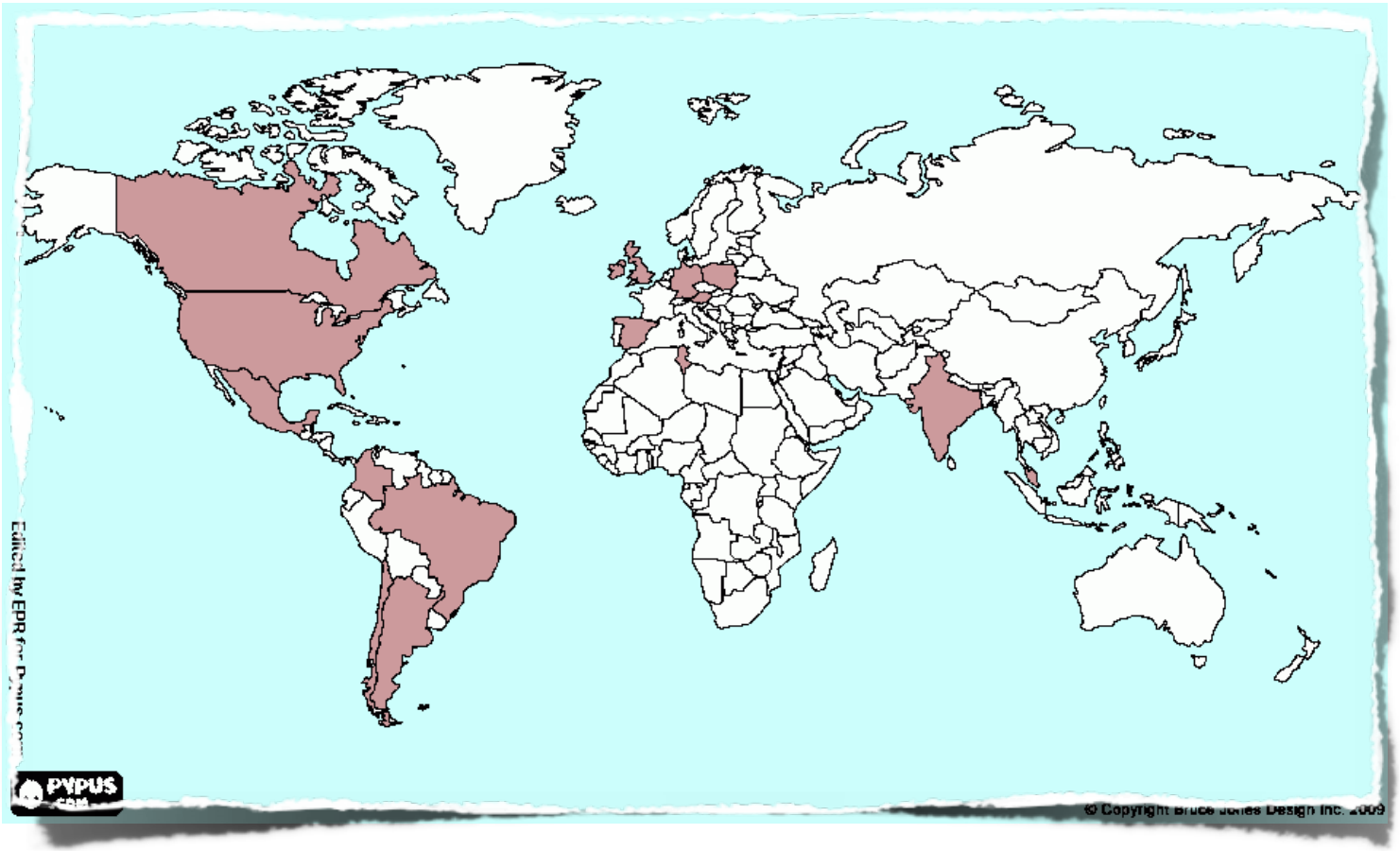
Other performance measures

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
- All participants improved in the final evaluation, mainly Aleman

Final results

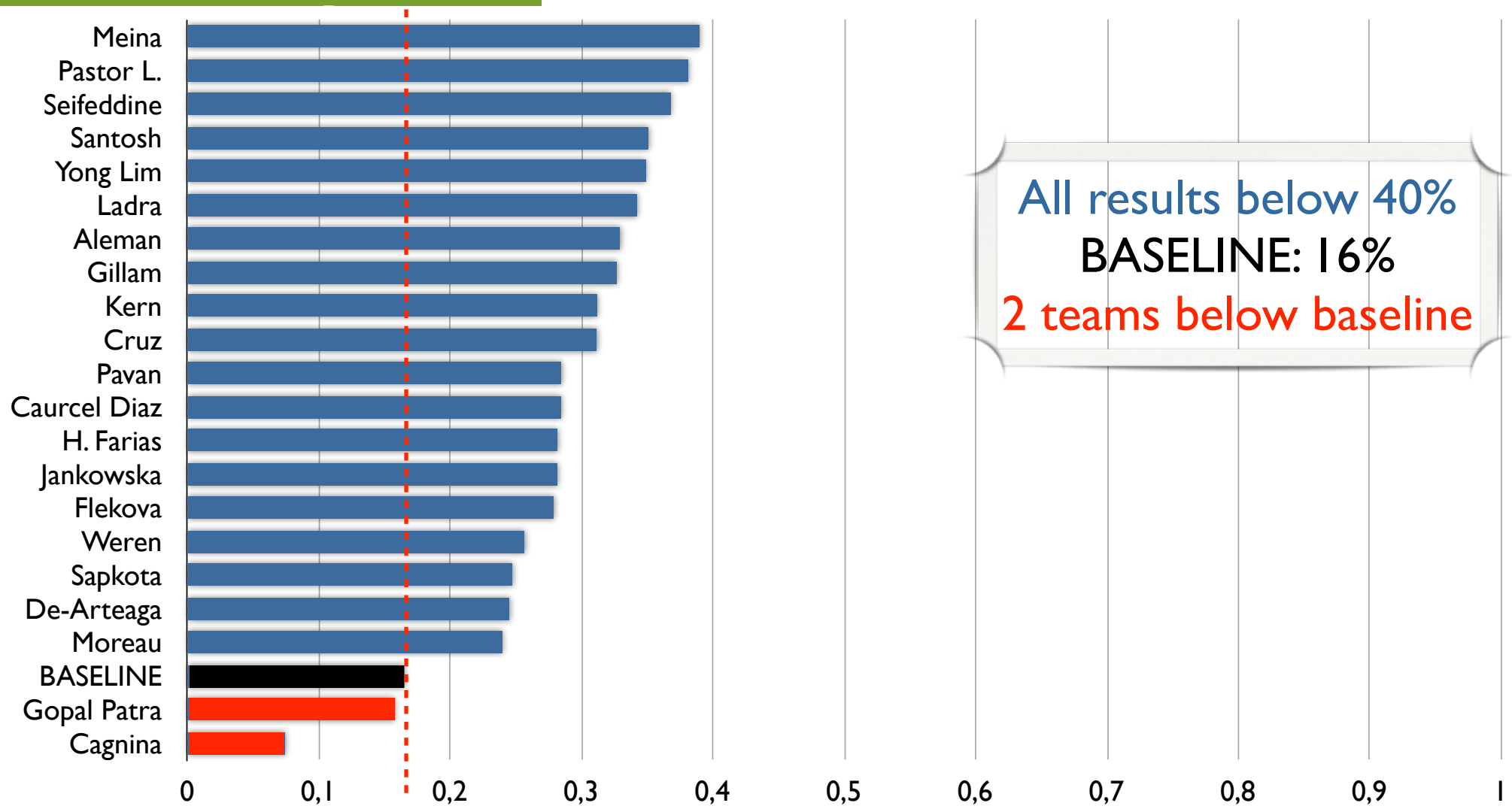
Table 3. Evaluation results in terms of accuracy on English (left) and Spansih (right) texts.

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

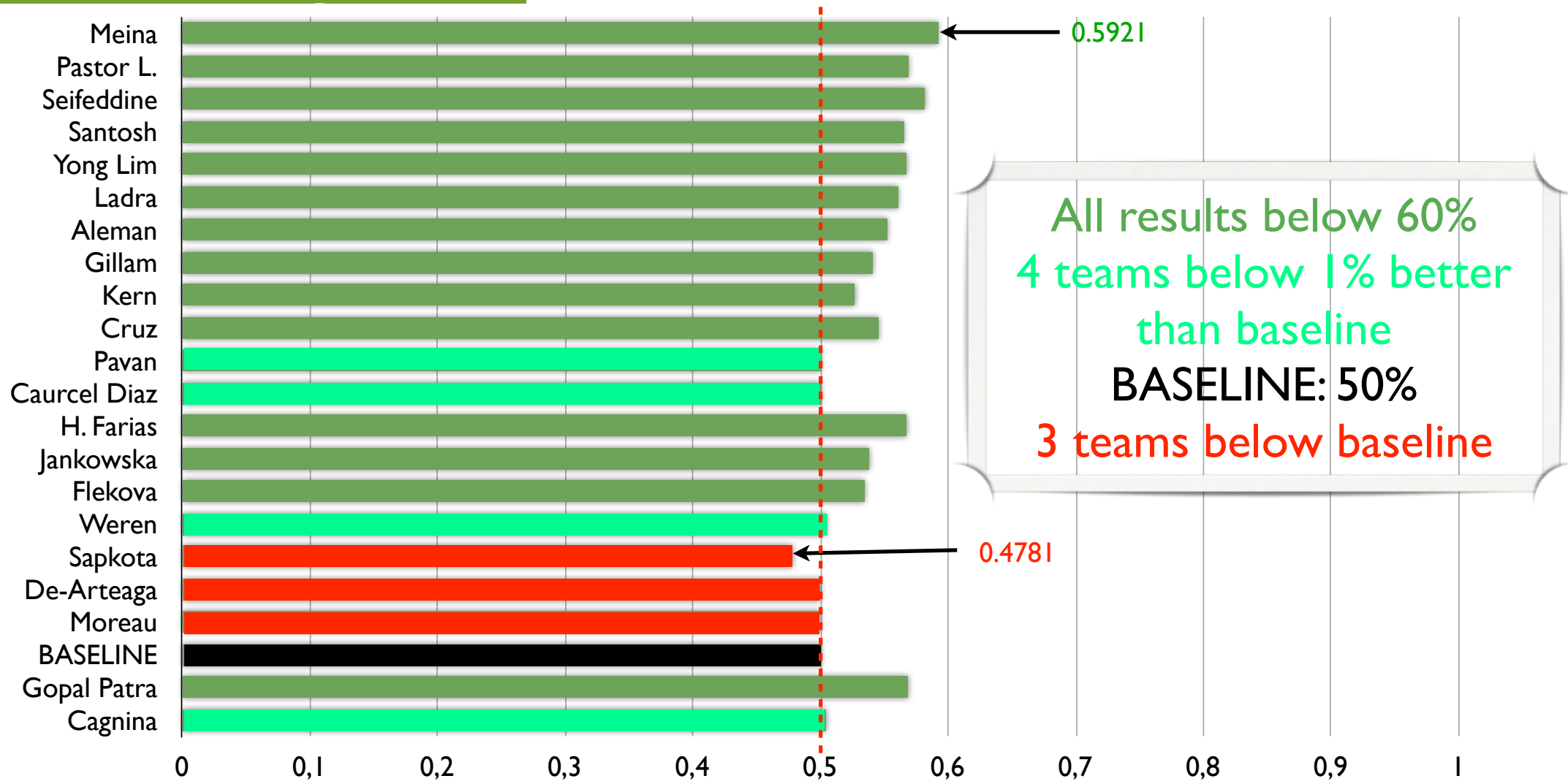
Spanish			
Team	Total	Gender	Age
Santosh	0.4208	0.6473	0.6430
Pastor L.	0.4158	0.6299	0.6558
Cruz	0.3897	0.6165	0.6219
Flekova	0.3683	0.6103	0.5966
Ladra	0.3523	0.6138	0.5727
De-Arteaga	0.3145	0.5627	0.5429
Kern	0.3134	0.5706	0.5375
Yong Lim	0.3120	0.5468	0.5705
Sapkota	0.2934	0.5116	0.5651
Pavan	0.2824	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



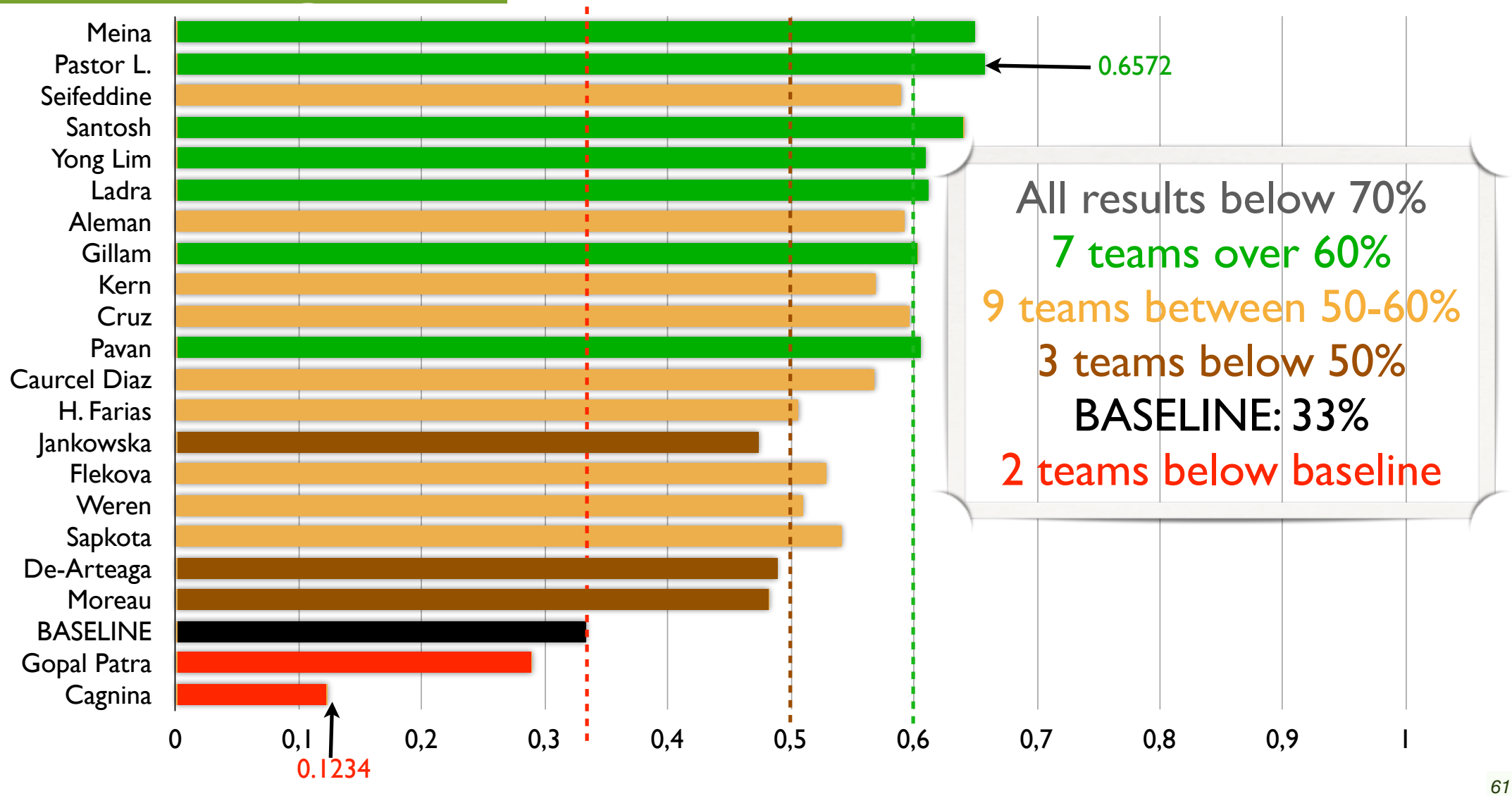
All results below 60%

4 teams below 1% better than baseline

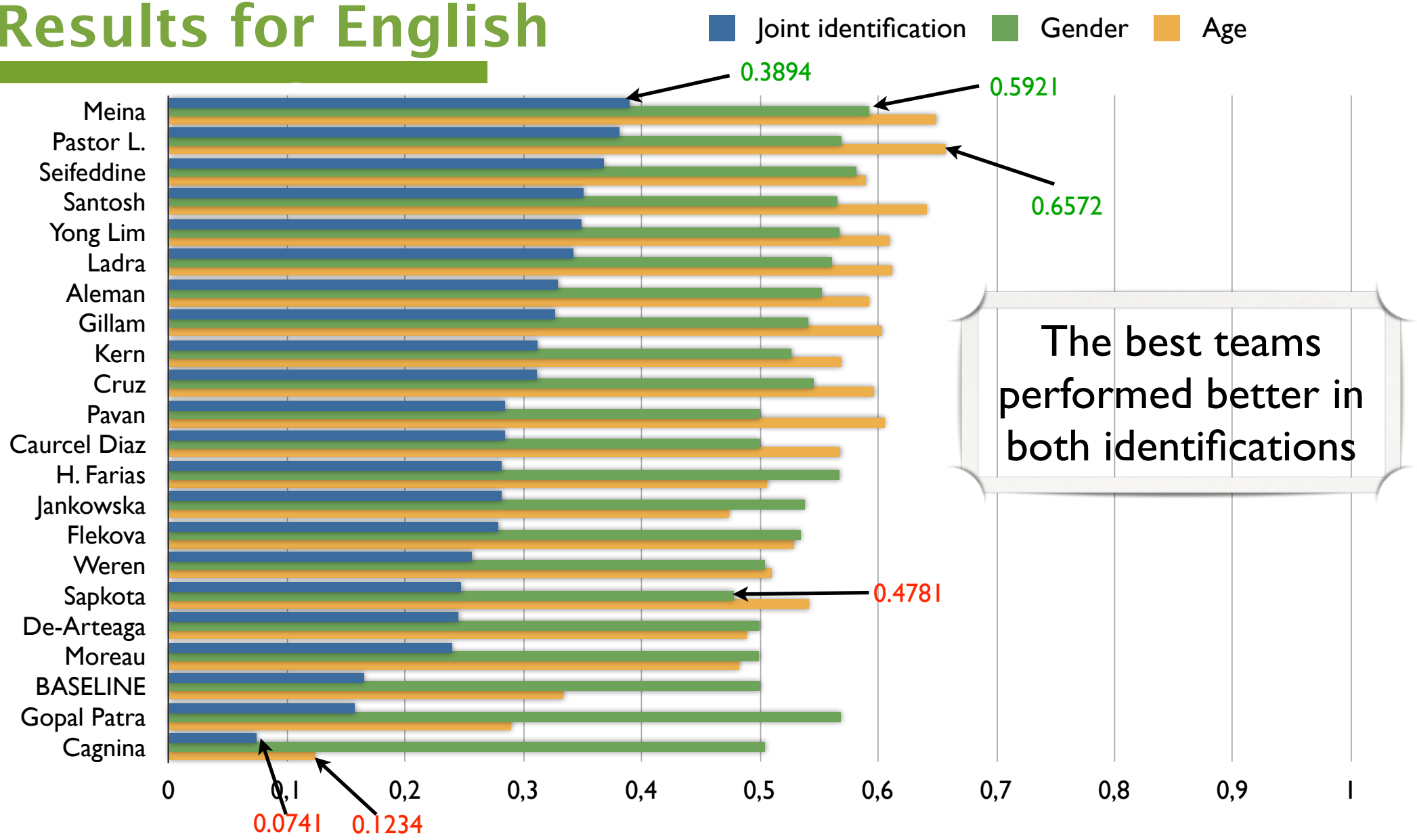
BASELINE: 50%

3 teams below baseline

Results for English

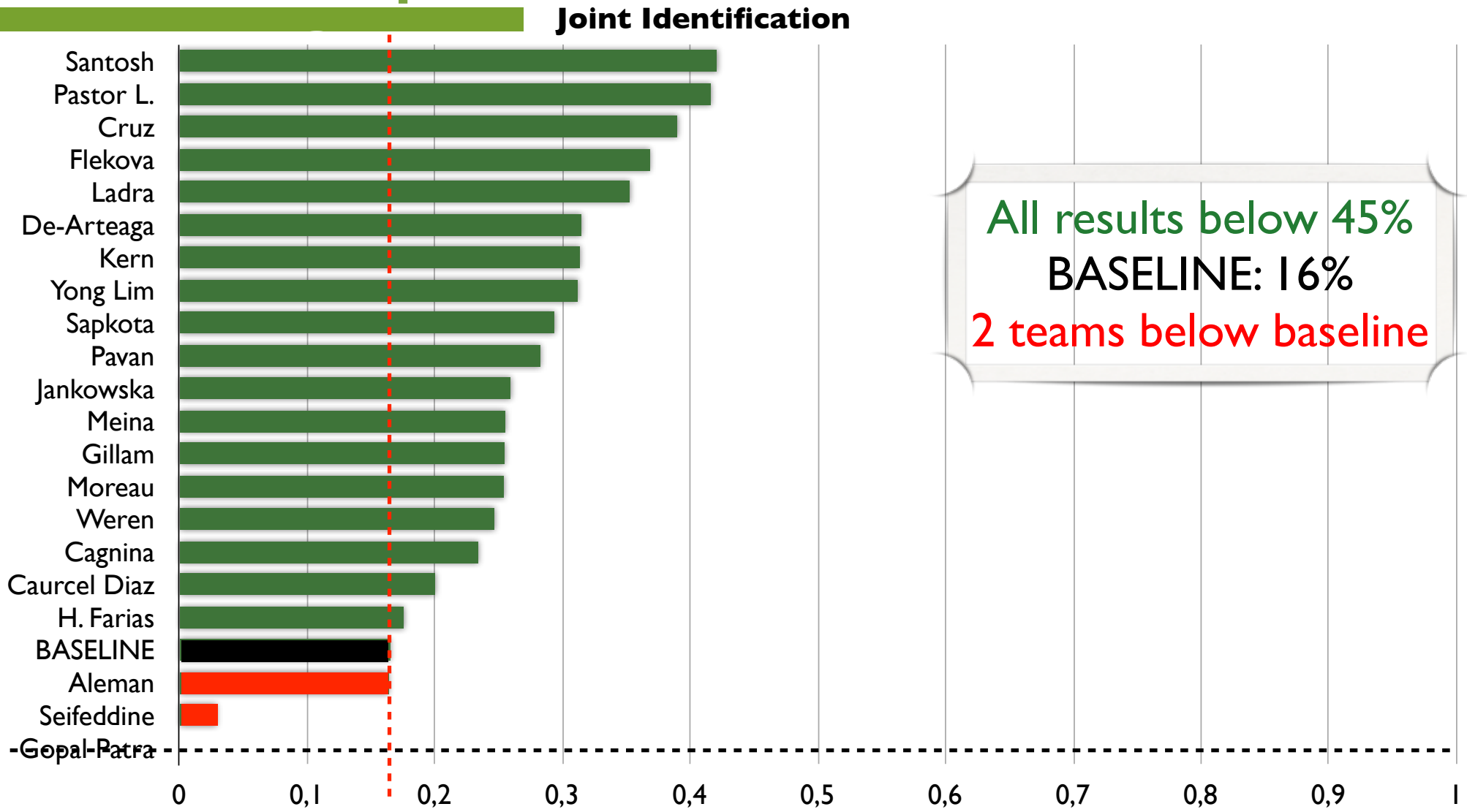


Results for English



The best teams performed better in both identifications

Results for Spanish

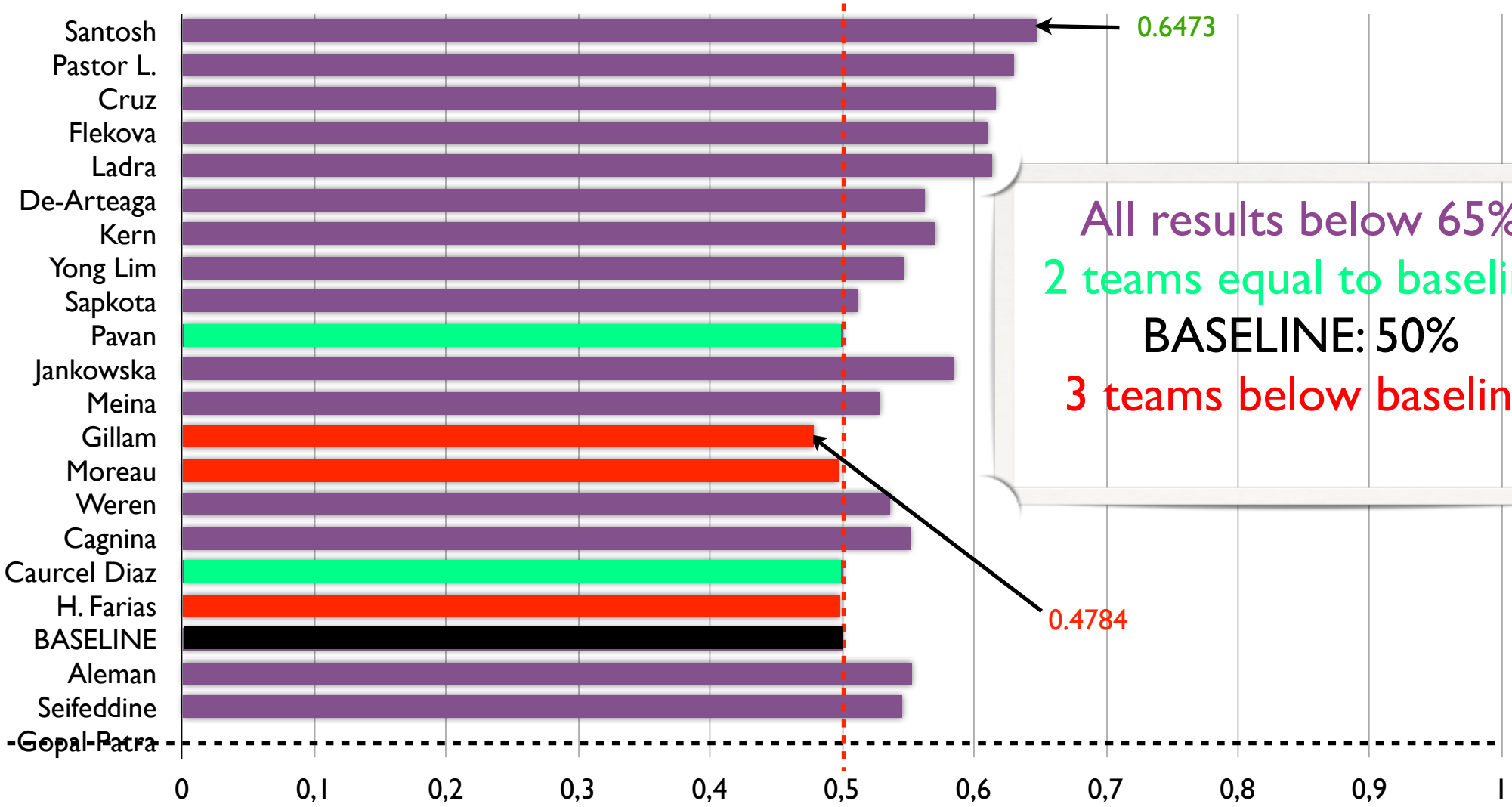


All results below 45%

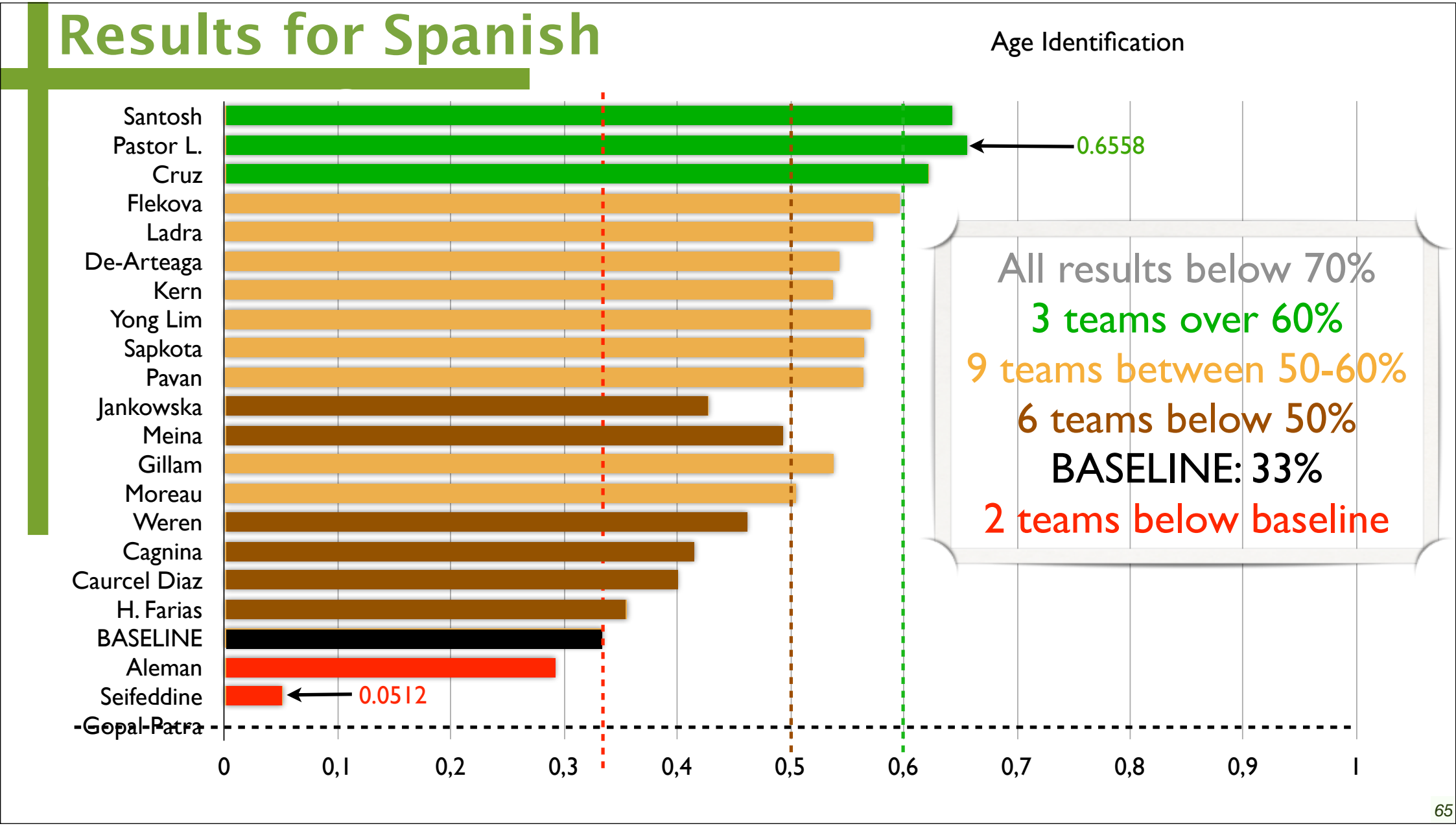
BASELINE: 16%

2 teams below baseline

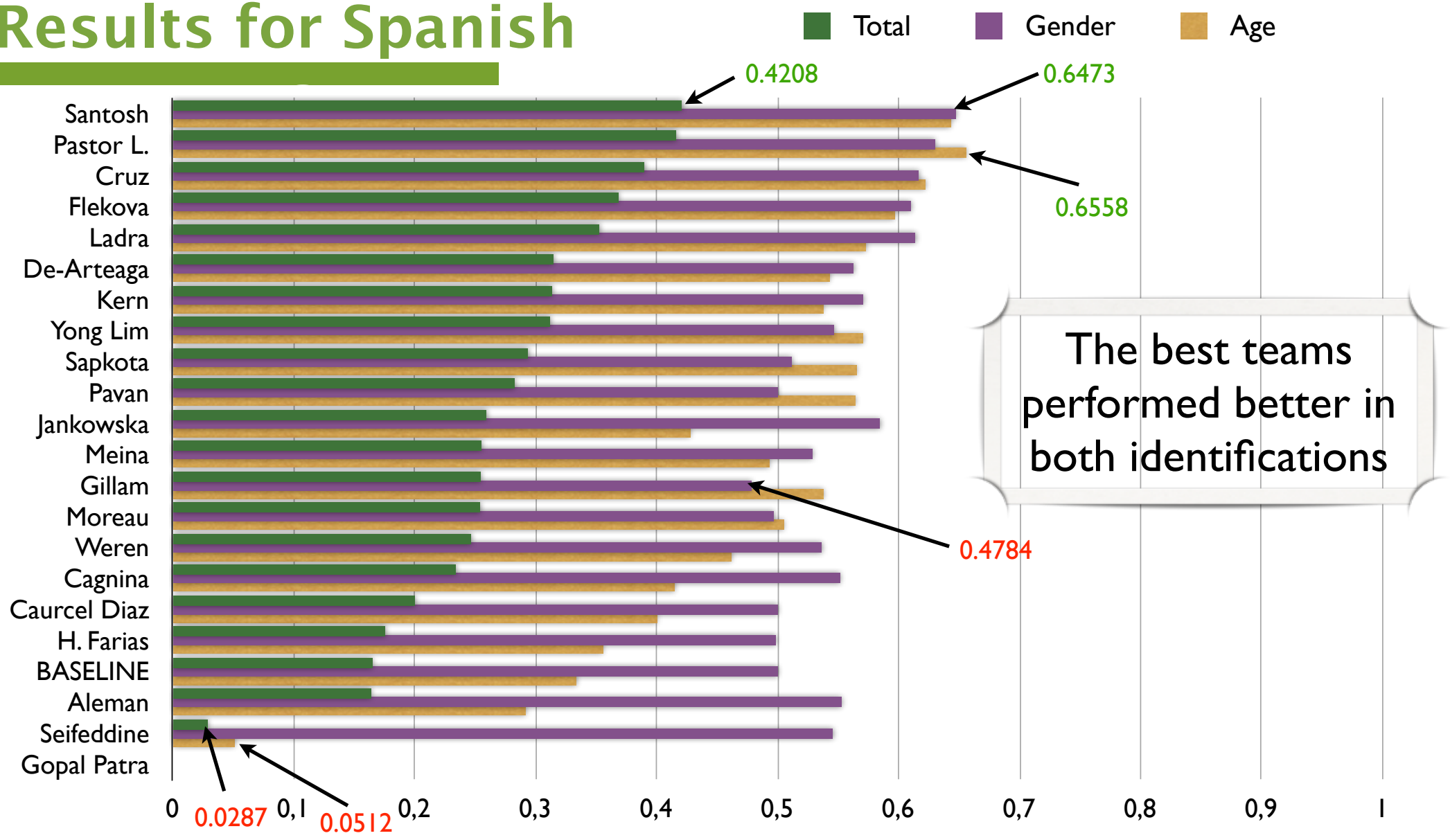
Results for Spanish



Results for Spanish

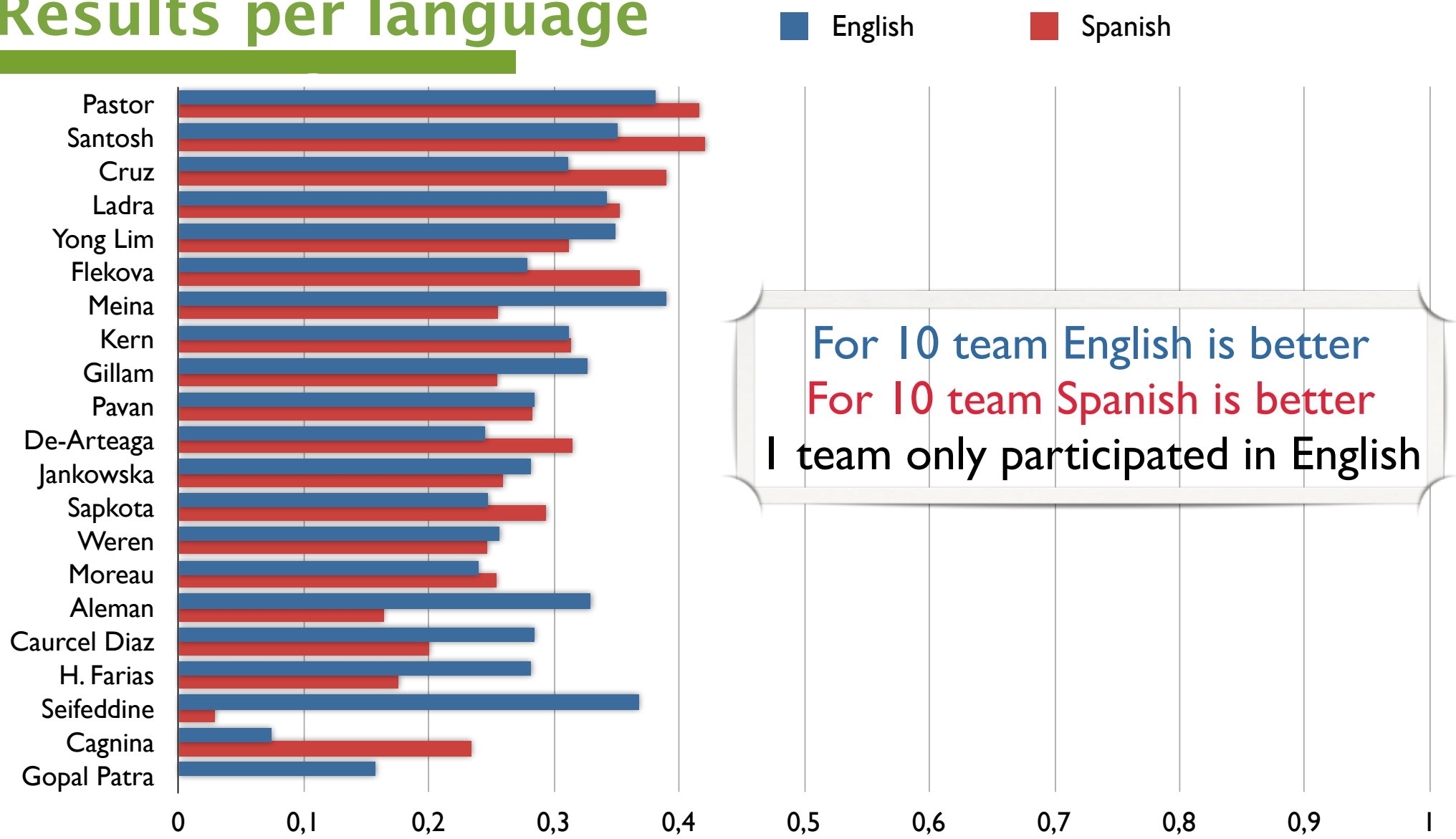


Results for Spanish



The best teams performed better in both identifications

Results per language



For 10 team English is better

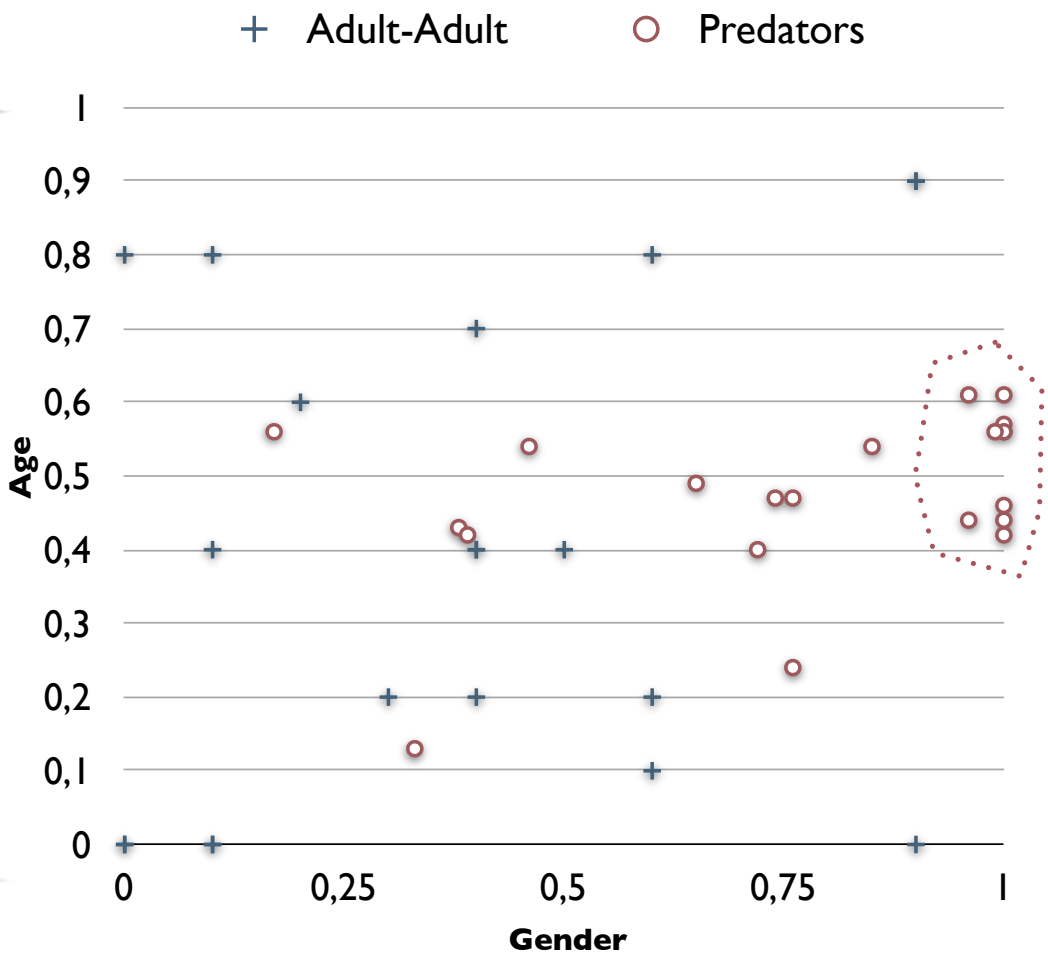
For 10 team Spanish is better

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?

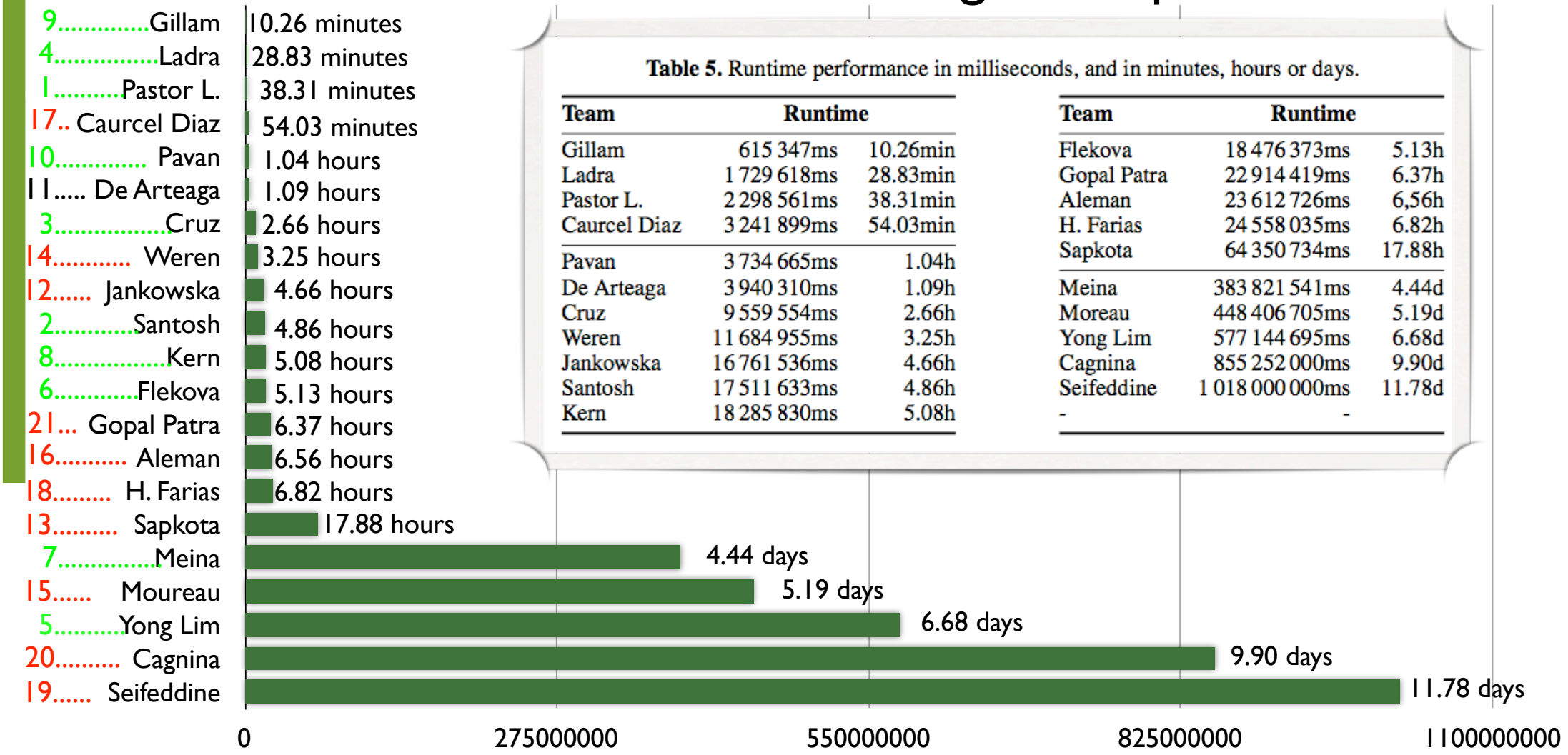


Table 5. Runtime performance in milliseconds, and in minutes, hours or days.

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 highest number of participants at CLEF (21)
- Interest from many teams (66 registered) but the task was new and many did not make it (potentially more participation next year!)

3. Personality Traits & Native Language Identification

Personality Recognition Shared task @ ICWSM 2013

http://mypersonality.org/wiki/doku.php?id=wcpr13

Shared task @ BEA8 2013

Native Language Identification:

https://sites.google.com/site/nlisharedtask2013/home

Personality Traits

- **Big Five** personality traits (given text, determine if author is):
  - **Open** (to new experiences)
  - **Conscientious** (tends to be careful and scrupuolus)
  - **Neurotic** (tends to worry about things)
  - **Extroverted** (gets energy from being around people)
  - **Agreeable** (prefers to agree with others)

- Students wrote essays and same students took personality assessment tests [J. W. Pennebaker]

Native Language Identification: Translationese

Given an English text, can we determine the **author's native language**?

Moshe Koppel
Bar-Ilan University

Exercise: Which is Which?

These were written by Russian, French and Spanish speakers, respectively.

In the second part of this outhor's novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are pejudments of small groups, such as homosexuals, inmigrants, aids diseaseds, etc. But "political correctness" has have positive and negative consecuences.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won "Oskar" prize.

Possible Clues

Patterns of native language are typically reflected in how other languages are spoken (Rado, 61, Corder, 81):

- **Word selection**
- **Syntax**
- **Spelling**

Orthographic Idiosyncrasies

- Repeated letter (e.g. *remmit* instead of *remit*)
- Double letter appears once (e.g. *comit* instead of *commit*)
- Letter α instead of β (e.g. *firsd* instead of *first*)
- Letter inversion (e.g. *fisrt* instead of *first*)
- Inserted letter (e.g. *friegnd* instead of *friend*)
- Missing letter (e.g. *frend* instead of *friend*)
- Conflated words (e.g *stucktogether*)

Syntactic Idiosyncrhasies

- Sentence Fragment
- Run-on Sentence
- Repeated Word
- Missing Word
- Mismatched Singular/Plural
- Mismatched Tense
- *that/which* confusion
- Rare POS pairs (Chodorow-Leacock, 00)

Automatically Finding Idiosynchrasies

1. Run text through automated spell/grammar checker
2. Compare flagged word to best suggestion
3. Mark error accordingly

e.g. text=*remmit* suggestion=*remit*

mark as "repeated letter"

Summary: Features Used

- 400 function words
- 200 letter sequences
- 185 error types
- 250 rare POS pairs

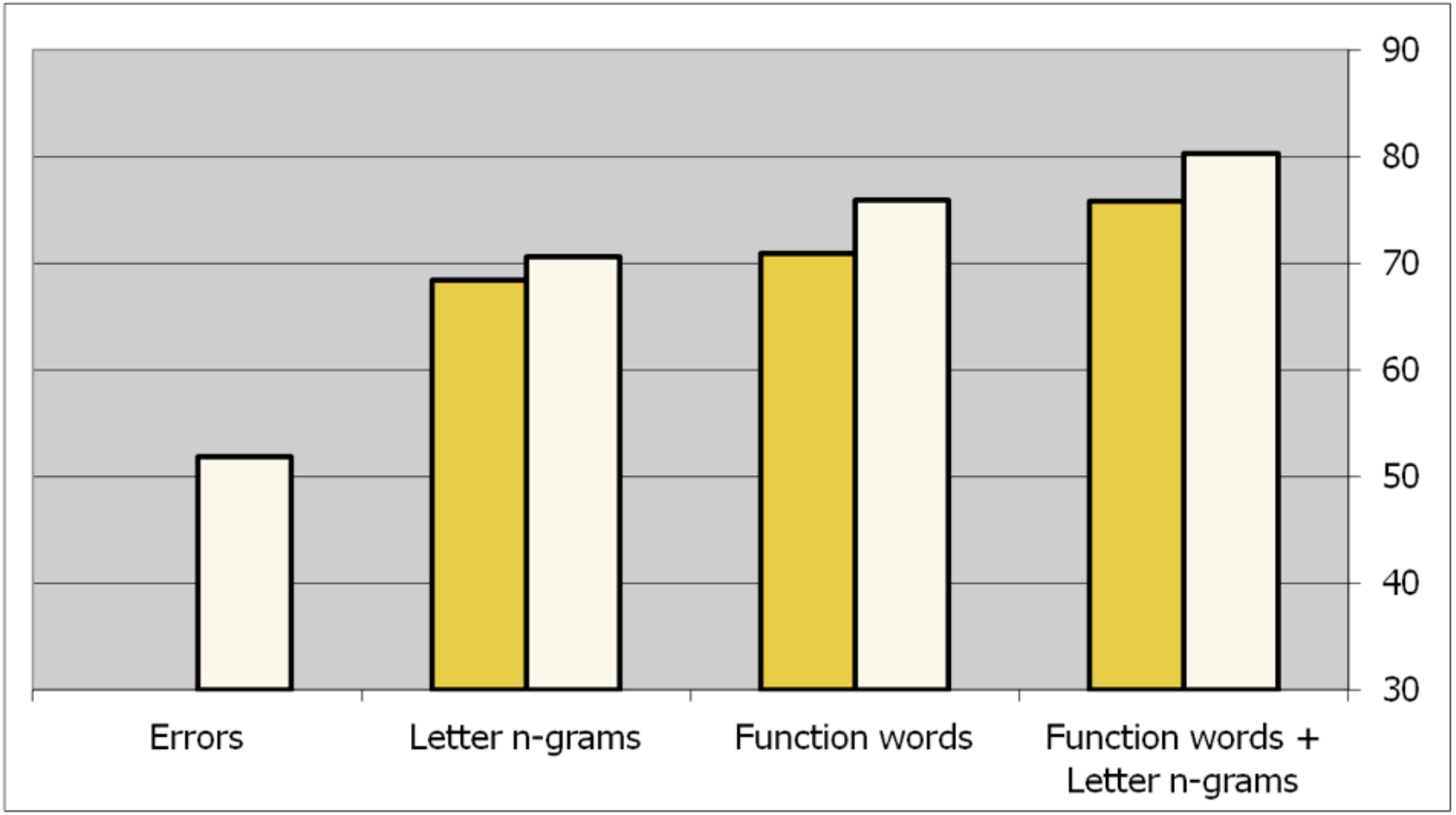
Each document is represented as numerical vector of length 1035

Test Corpus

International Corpus of Learner English (Granger, 98)

- 11 countries
- Subjects same age, proficiency level
- Samples same genre, length
- Actually used in study- 258 docs:
  - French
  - Spanish
  - Bulgarian
  - Czech
  - Russian

Classification Accuracy (10-fold CV)



Baseline=20% (5 languages)

Confusion Matrix

Actual \ Classified as	Czech	French	Bulgarian	Russian	Spanish
Czech	**209**	1	18	20	10
French	9	**219**	13	12	5
Bulgarian	14	8	**211**	18	7
Russian	24	8	24	**194**	8
Spanish	16	10	10	7	**215**

What Gives it Away?

- Russian –*over*, *the* (infrequent), **number_relative-adverb**
- French – *indeed, Mr* (no period), misused *o (*e.g. ***o****uthor)*
- Spanish – *c-q* confusion (e.g. ***c****uality*), *m-n* confusion (e.g. *co****n****fortable*), undoubled consonant (e.g. *co****m****it*)
- Bulgarian – *most*_**adverb**, *cannot* (uncontracted)
- Czech – doubled consonant (e.g. *re****mm****it*)

Exercise: Play it again, Sam...

In the second part of this outhor's novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are pejudments of small groups, such as homosexuals, inmigrants, aids diseaseds, etc. But "political correctness" has have positive and negative consecuences.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won "Oskar" prize.

Exercise: Now it's pretty obvious!

In the second part of this outhor's novel, called Time Passes, time has passed indeed and Mrs Ramsay has died.

There are pejudments of small groups, such as homosexuals, inmigrants, aids diseaseds, etc. But "political correctness" has have positive and negative consecuences.

There is one more kind of films irritating many television viewers - "soap" serials. «Santa Barbara» has even won [the] "Oskar" prize.

Real-Life Difficulties

- Many candidate languages
- Very short texts
- Unpredictable English proficiency

Thanks!!

Paolo Rosso
Universitat Politècnica de València
prosso@dsic.upv.es
http://www.upv.es/~prosso

I hope now you'll be able to profile who ruins your reputation!
And don't hesitate in participating in the Author Profiling task @ PAN next year ;-)