UNIVERSITY OF SURREY

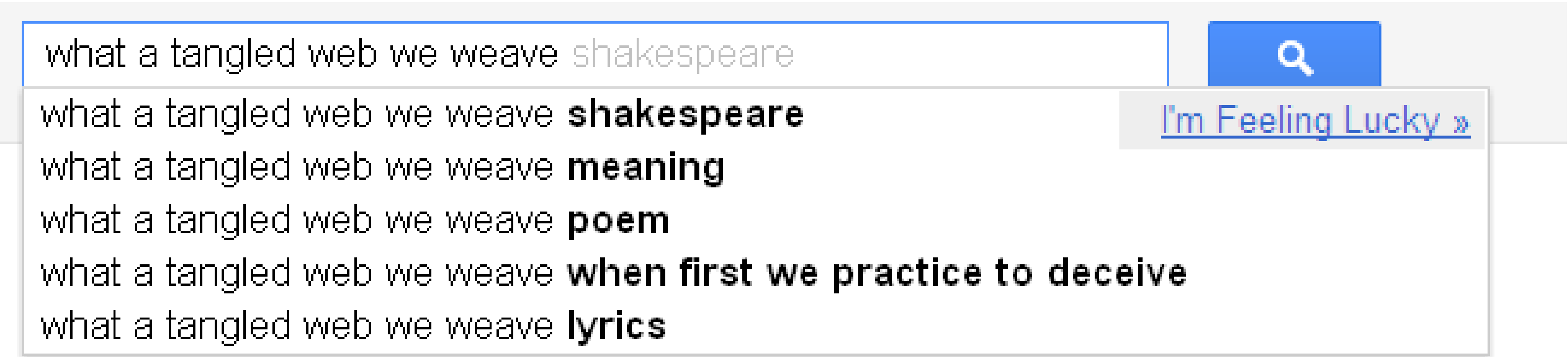
Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification

Notebook for PAN at CLEF 2012

Anna Vartapetiance

Dr. Lee Gillam

“Oh what a tangled web we weave,
When first we practice to deceive”



what a tangled web we weave shakespeare

what a tangled web we weave **shakespeare**

what a tangled web we weave **meaning**

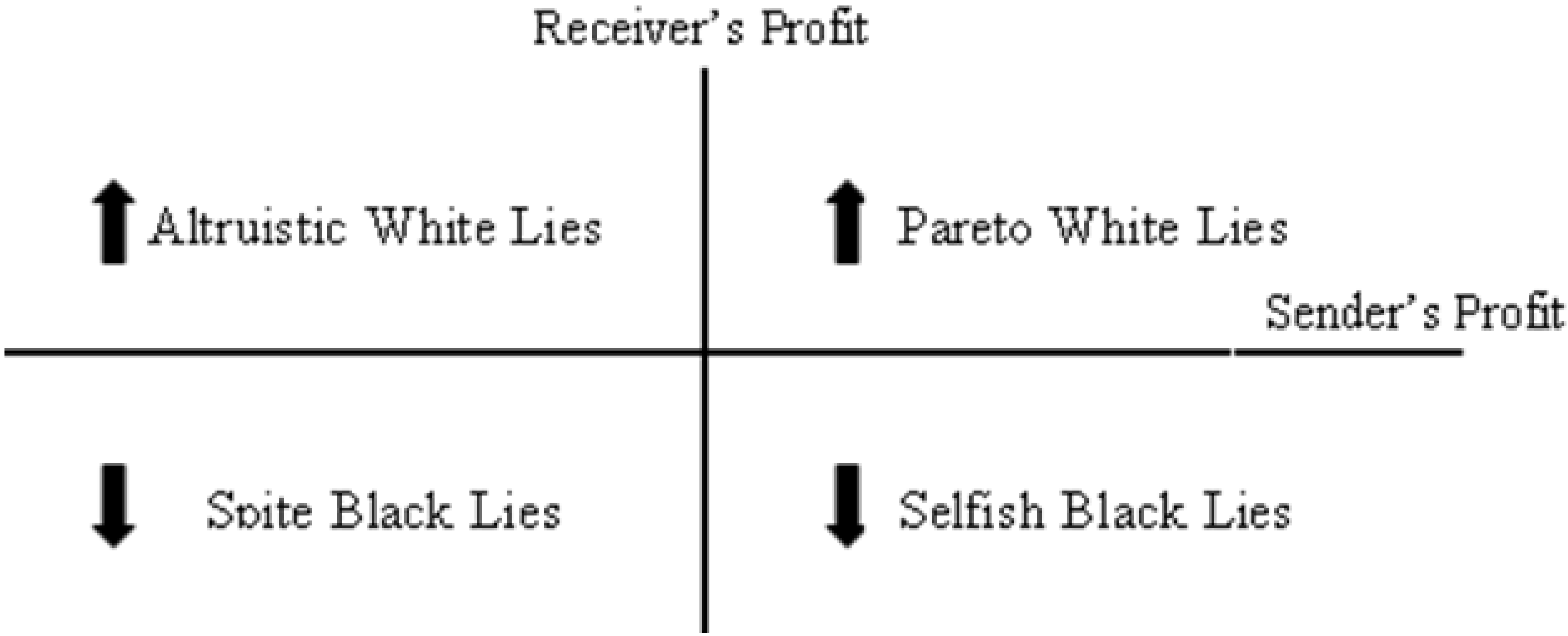
what a tangled web we weave **poem**

what a tangled web we weave **when first we practice to deceive**

what a tangled web we weave **lyrics**

Magnitude of Deception and Acceptability

- Classification of Deception (Lies) : Magnitude
  - Based on their level a acceptance
  - Erat and Gneezy (2009)



Vartapetiance, A., Gillam, L.: "I don't know where he's not": Does Deception Research yet offer a basis for Deception Detectives?: Proceedings of the Workshop on Computational Approaches to Deception Detection, pp. 3-14, Avignon, France (2012)

Deception Detection

- Deception Cues → 3Vs
  - Visual
  - Vocal
  - Verbal *****

- What can flag Verbal Deception ?
  - Quantity: e.g. word count, average of words per sentence
  - Quality: lexical selections, e.g. number of verbs and nouns
  - Overall impression: human judgement, e.g. sounding helpful

- What is out there? And why it is not working
  - Generalized Cues: DePaulo et al. (2003) → 158 cues, 25 measurable
  - Frequency-based Cues: Pennebaker → self-references, negative words, Exclusive words, Action verbs
  - Category-based Cues: Burgoon →45 cues in 8 categories but inconsistent in both categories and membership

Categories	Cues	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
Quantity	Syllables	--	--	--	--	**		--	--	--	--	--	--	-**	Q
	Word	**	Q	**	Q	**		+**	Q	+**	Q	-**	Q	-**	Q
	Sentence	**	Q	**	Q	**		+**	Q	+**	Q	-**	Q	-**	Q
	Noun phrase	--	--	--	--	--		+**	Q	+**	Q	**	Q	--	--
Specificity	Sensory details	**	S	**	S	**		-***	--	--	--	--	--	--	--
	Modifiers	**	S	-**	U	--		+**	U	**	Q	**	Q	--	--
	First-person singular	**	S	--	--	--		-**	V	+**	V	-**	V	--	--
	2nd person pronouns	**	S	--	--	--		-**	U	**	V	--	--	--	--
	3rd person pronouns	**	S	--	--	--						**	V	--	--
	Temporal details	--	--	**	S	--		+**	S	--	--	-**	S	--	--
	Spatial details	--	--	**	S	--				--	--			--	--
	Over all specificity	--	--	**	S	--				--	--			--	--
	Perceptual information	--	--	--	--	--		+**	S	--	--	-**	S	--	--
Affect	Affective terms	**	A	**	A	**		--	--	--	--	--	--	-**	S
	Imagery	**	A	**	A	--		--	--	--	--	--	--	--	--
	Positive	--	--	--	--	--		+**	S	+**	A	-**	S	--	--
	Negative	--	--	--	--	--		+**	S	+**	A	+**	S	--	--
Activation / Expressiveness	Emotiveness index	**	E	--	--	**		+**	E	--	--	+**	E	-**	S
	Activation	**	E	--	--	--		--	--	--	--	--	--	--	--
Diversity	Lexical diversity	**	D	-**	D	--		-**	D	-**	D	-**	D	--	--
	Content word diversity	**	D	**	D	--		-**	D	-**	D	-**	D	--	--
	Redundancy	**	D	-**	D	--		-**	D	**	D	+**	D	--	--
Verbal non-immediacy	Passive voice	**	V	**	V	--		+ **	V	**	V	+ **	V	--	--
	Reference	--	--	**	V	--		--	--	--	--	--	--	--	--
	modal verbs	**	U	-**	U	--		+**	U	**	V	+**	V	--	--
	Uncertainty	--	--	+***	--	--		-**	U	**	V	+**	V	--	--
	Objectification	--	--	--	--	--		-**	V	+**	V	**	V	--	--
	Generalising term	--	--	--	--	--		-**	V	-**	V	**	V	--	--
Informality	Typo errors	--	--	-***	--	**		+**	I	+**	I	+**	I	--	--

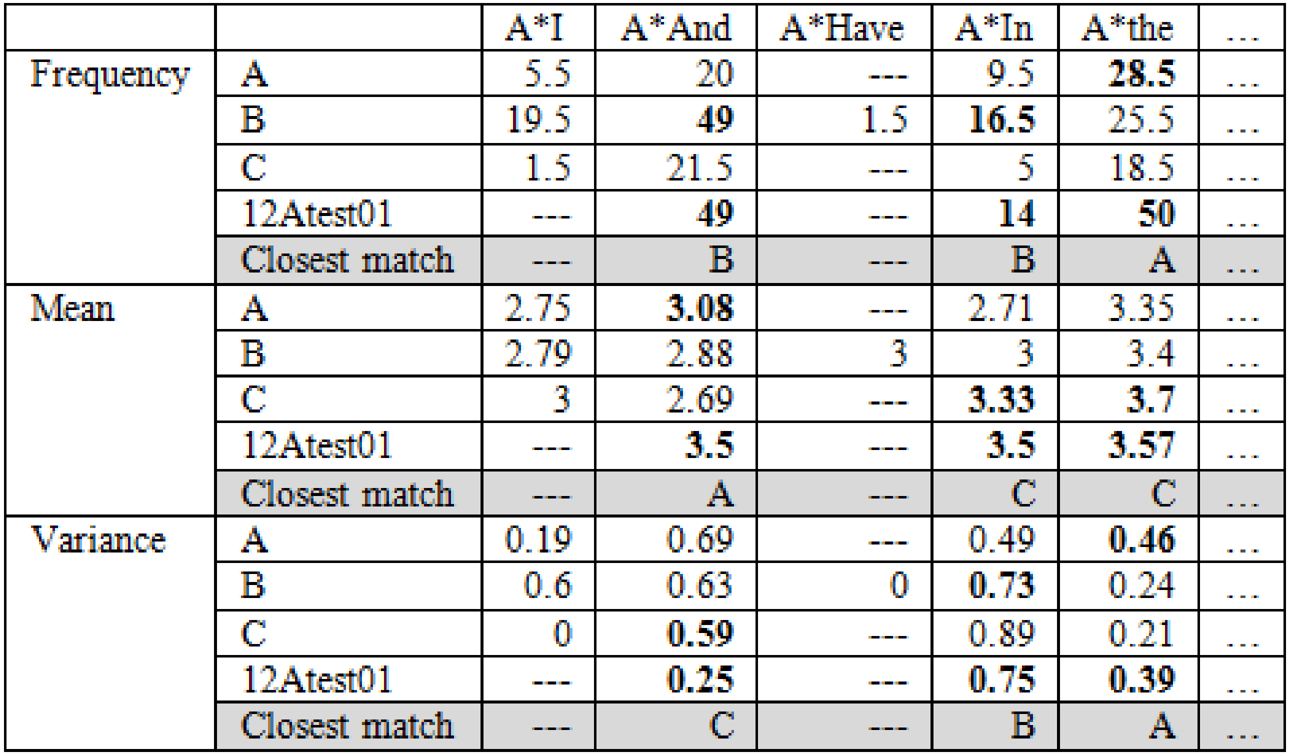
Quantity = Q; Complexity = C; Specificity = S; Affect = A; Activation /Expressiveness = E; Diversity = D; Verbal non-immediacy = V; Informality = I; Uncertainty = U; Vocabulary Complexity = VC; Grammatical Complexity = GC;
(1) Burgoon & Qin, 2006 (2) Qin et al. 2005 (3) Qin, Burgoon & Nunamaker, 2004 (4) Zhou et al. 2004 (5) Zhou, Burgoon & Twitchell, 2003 (6) Zhou et al. 2003 (7) Burgoon et al. 2003

Authorship Attribution: Closed dataset

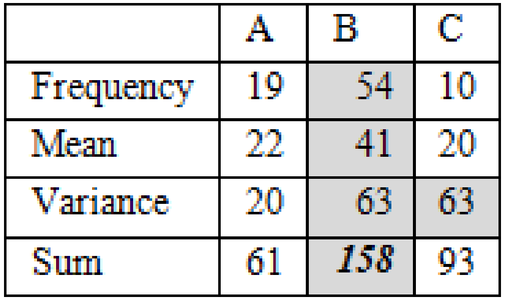
1. Top 10 most frequent words (English)
   - the, be, to, of, and, a, in, that, have, I
2. Regular expressions for all paired, with specific window size
   - the + have, have + the
   - window size of 5
3. Create author profiles based on the patterns
4. Calculate frequency, mean, variance of the patterns for each author (mean-variance, following Church & Hanks, 1991)
5. Calculate frequency, mean and variance for each test document
6. Select the author with closest match values

Church, K., and Hanks, P. (1991). Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, Vol 16:1, pp. 22-29

Authorship Attribution: Closed dataset



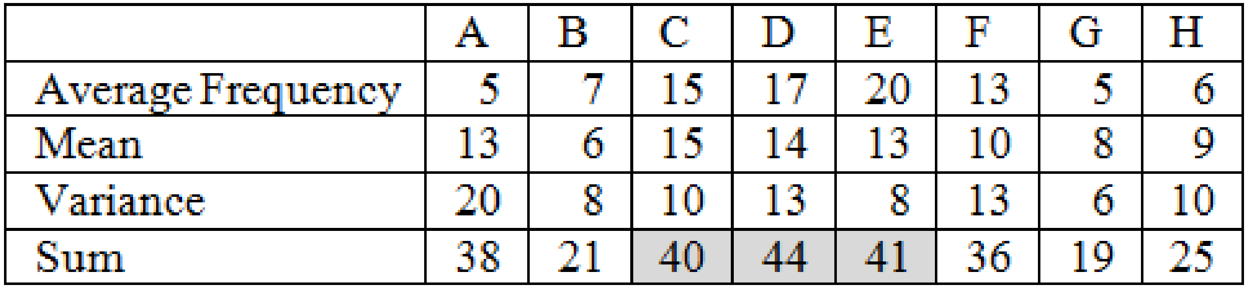
		A*I	A*And	A*Have	A*In	A*the	…
Frequency	A	5.5	20	---	9.5	**28.5**	…
	B	19.5	**49**	1.5	**16.5**	25.5	…
	C	1.5	21.5	---	5	18.5	…
	12Atest01	---	**49**	---	**14**	**50**	…
	Closest match	---	B	---	B	A	…
Mean	A	2.75	**3.08**	---	2.71	3.35	…
	B	2.79	2.88	3	3	3.4	…
	C	3	2.69	---	**3.33**	**3.7**	…
	12Atest01	---	**3.5**	---	**3.5**	**3.57**	…
	Closest match	---	A	---	C	C	…
Variance	A	0.19	0.69	---	0.49	**0.46**	…
	B	0.6	0.63	0	**0.73**	0.24	…
	C	0	**0.59**	---	0.89	0.21	…
	12Atest01	---	**0.25**	---	**0.75**	**0.39**	…
	Closest match	---	C	---	B	A	…



	A	B	C
Frequency	19	54	10
Mean	22	41	20
Variance	20	63	63
Sum	61	***158***	93

Authorship Attribution: Open dataset

- Special Condition over “NA”
  - If difference between 1st and 2nd highest value is less than 5, “NA”
  - Else select the highest match

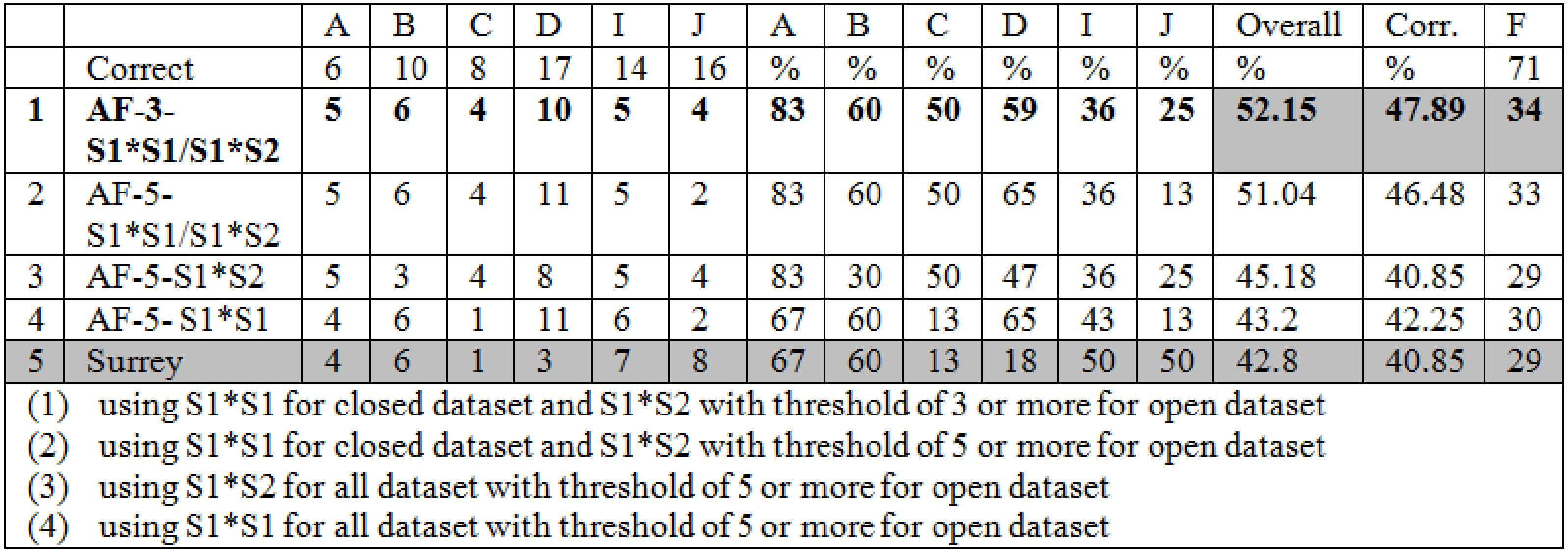


	A	B	C	D	E	F	G	H
Average Frequency	5	7	15	17	20	13	5	6
Mean	13	6	15	14	13	10	8	9
Variance	20	8	10	13	8	13	6	10
Sum	38	21	40	44	41	36	19	25

- Results
  - 40.85% (29 out of 71)

Improvements?

- Post-competition analysis
  - Vary window size (5, 10 and 25)
  - Vary confidence for Open dataset (2,3,5 and 10)
  - Vary numbers of stopwords (5*5)
- Best results: S1*S1 for closed and S1*S2 for Open datasets
  - S1: the, be, to, of, and
  - S2: a, in, that, have, I



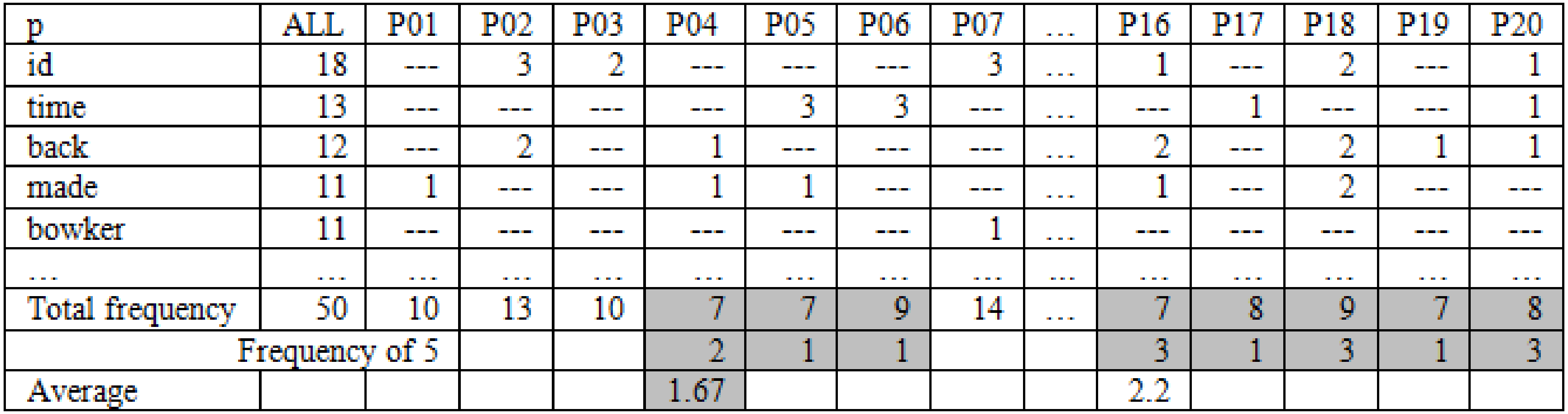
		A	B	C	D	I	J	A	B	C	D	I	J	Overall	Corr.	F
	Correct	6	10	8	17	14	16	%	%	%	%	%	%	%	%	71
1	**AF-3-S1*S1/S1*S2**	**5**	**6**	**4**	**10**	**5**	**4**	**83**	**60**	**50**	**59**	**36**	**25**	**52.15**	**47.89**	**34**
2	AF-5-S1*S1/S1*S2	5	6	4	11	5	2	83	60	50	65	36	13	51.04	46.48	33
3	AF-5-S1*S2	5	3	4	8	5	4	83	30	50	47	36	25	45.18	40.85	29
4	AF-5- S1*S1	4	6	1	11	6	2	67	60	13	65	43	13	43.2	42.25	30
5	Surrey	4	6	1	3	7	8	67	60	13	18	50	50	42.8	40.85	29

(1) using S1*S1 for closed dataset and S1*S2 with threshold of 3 or more for open dataset
(2) using S1*S1 for closed dataset and S1*S2 with threshold of 5 or more for open dataset
(3) using S1*S2 for all dataset with threshold of 5 or more for open dataset
(4) using S1*S1 for all dataset with threshold of 5 or more for open dataset

Intrinsic Plagiarism: Task F

1. 50 most frequent words for each file after removing stopwords
2. Determining frequency by paragraphs for these 50 words
3. Selecting (sequences of) paragraphs with fewer similarities (10)

- If there is more than one candidate sequence then select the longest sequence of paragraphs that
  - Does not share the most frequent words and
  - Has the highest average frequency for top 5 words

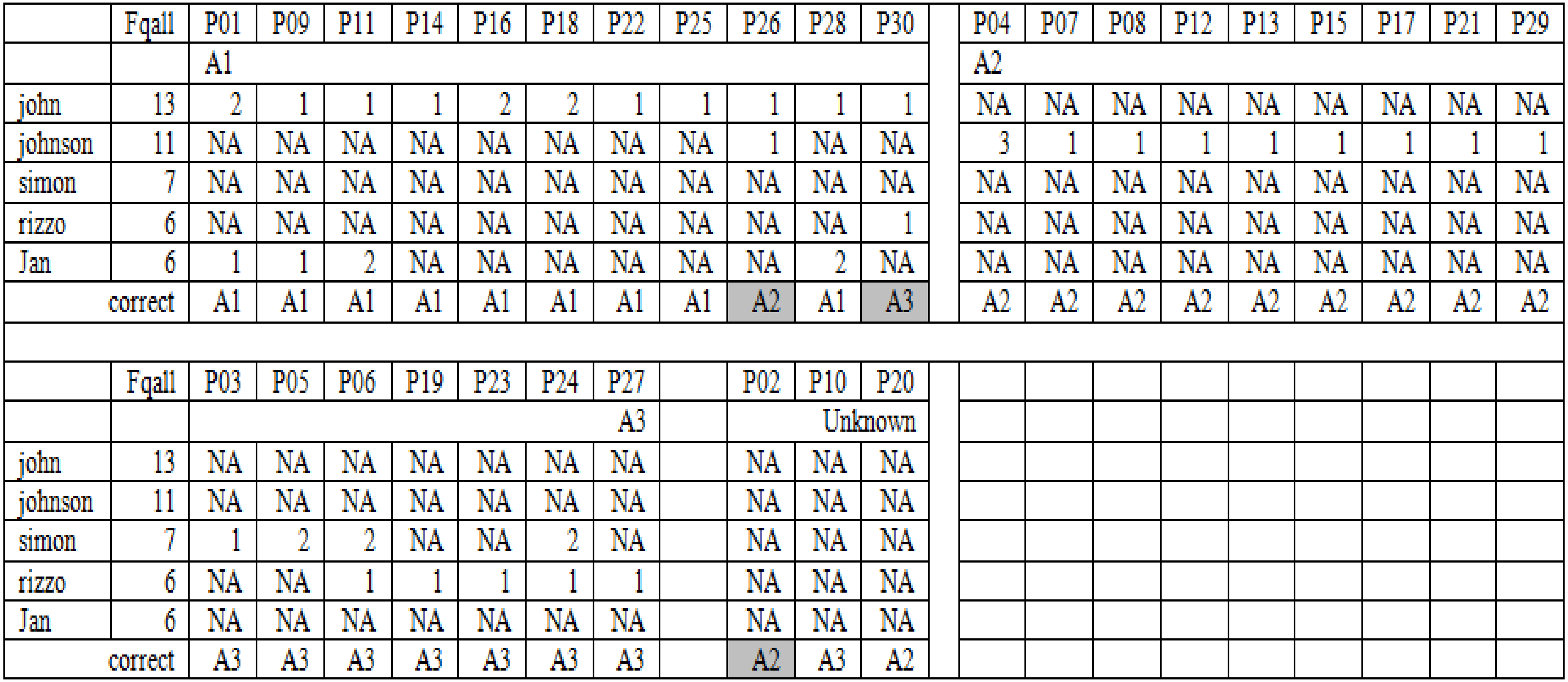


p	ALL	P01	P02	P03	P04	P05	P06	P07	…	P16	P17	P18	P19	P20
id	18	---	3	2	---	---	---	3	…	1	---	2	---	1
time	13	---	---	---	---	3	3	---	…	---	1	---	---	1
back	12	---	2	---	1	---	---	---	…	2	---	2	1	1
made	11	1	---	---	1	1	---	---	…	1	---	2	---	---
bowker	11	---	---	---	---	---	---	1	…	---	---	---	---	---
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
Total frequency	50	10	13	10	7	7	9	14	…	7	8	9	7	8
Frequency of 5					2	1	1			3	1	3	1	3
Average					1.67					2.2				

Intrinsic Plagiarism: Task E

1. Step 1 and 2 as Task F
2. Select proper nouns from the top 50
3. Create a cluster and remove from consideration all other linked nouns
4. Where the paragraphs are not allocated
   - If number of consecutive unallocated paragraphs > 5, then create a new cluster
   - Else, (a) paragraphs between two in the same cluster are allocated to the same cluster, (b) paragraphs between different clusters are allocated to the subsequent cluster
5. Results
   - Task F: 100% correct
   - Task E: 82.2% correct
   - 2nd in just this task (91.1% against 94.2%)

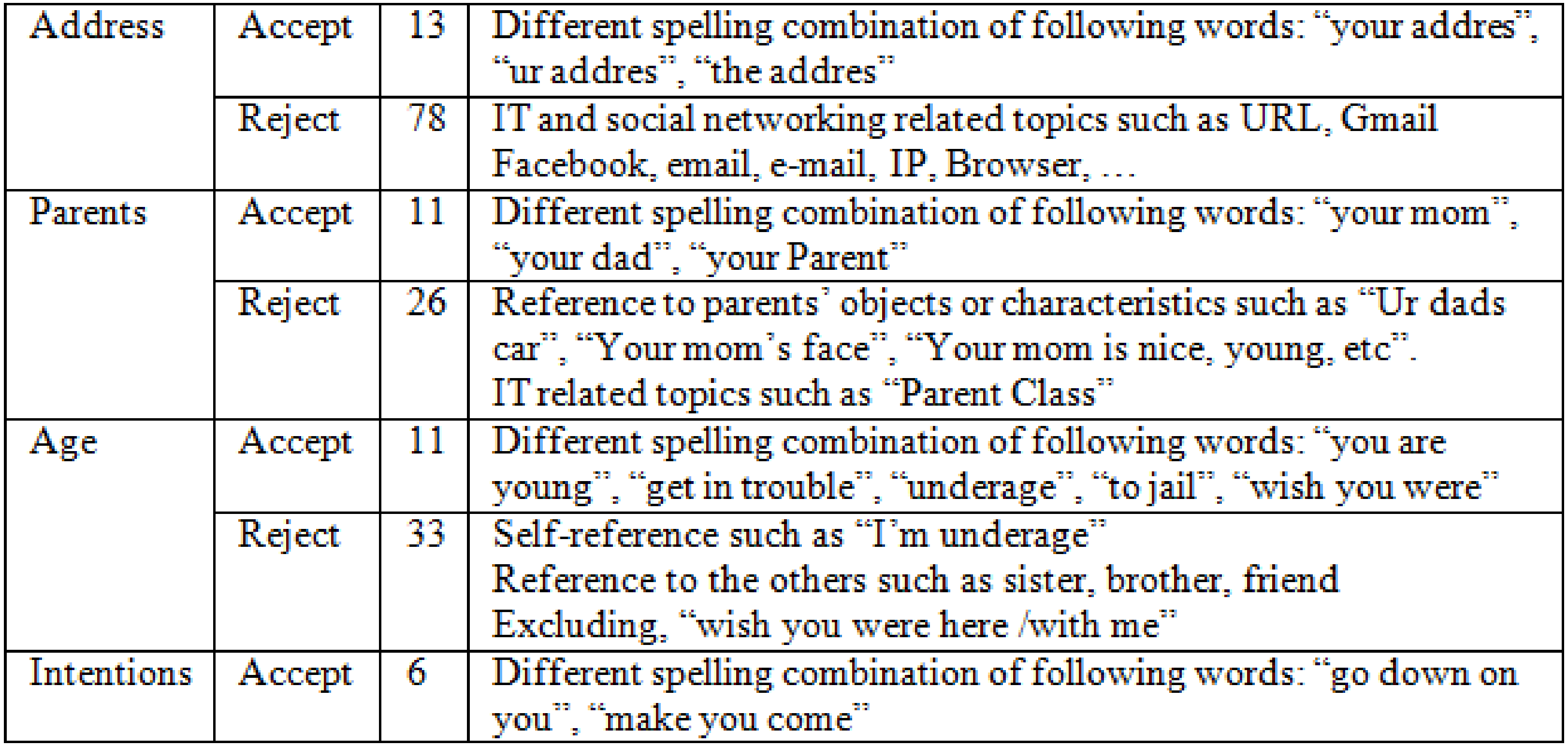
Intrinsic Plagiarism: Task E



	Fqall	P01	P09	P11	P14	P16	P18	P22	P25	P26	P28	P30		P04	P07	P08	P12	P13	P15	P17	P21	P29
		A1												A2								
john	13	2	1	1	1	2	2	1	1	1	1	1		NA	NA	NA	NA	NA	NA	NA	NA	NA
johnson	11	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA		3	1	1	1	1	1	1	1	1
simon	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA
rizzo	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1		NA	NA	NA	NA	NA	NA	NA	NA	NA
Jan	6	1	1	2	NA	NA	NA	NA	NA	NA	2	NA		NA	NA	NA	NA	NA	NA	NA	NA	NA
correct		A1	A1	A1	A1	A1	A1	A1	A1	A2	A1	A3		A2	A2	A2	A2	A2	A2	A2	A2	A2
	Fqall	P03	P05	P06	P19	P23	P24	P27		P02	P10	P20										
		A3								Unknown												
john	13	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA										
johnson	11	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA										
simon	7	1	2	2	NA	NA	2	NA		NA	NA	NA										
rizzo	6	NA	NA	1	1	1	1	1		NA	NA	NA										
Jan	6	NA	NA	NA	NA	NA	NA	NA		NA	NA	NA										
correct		A3	A3	A3	A3	A3	A3	A3		A2	A3	A2										

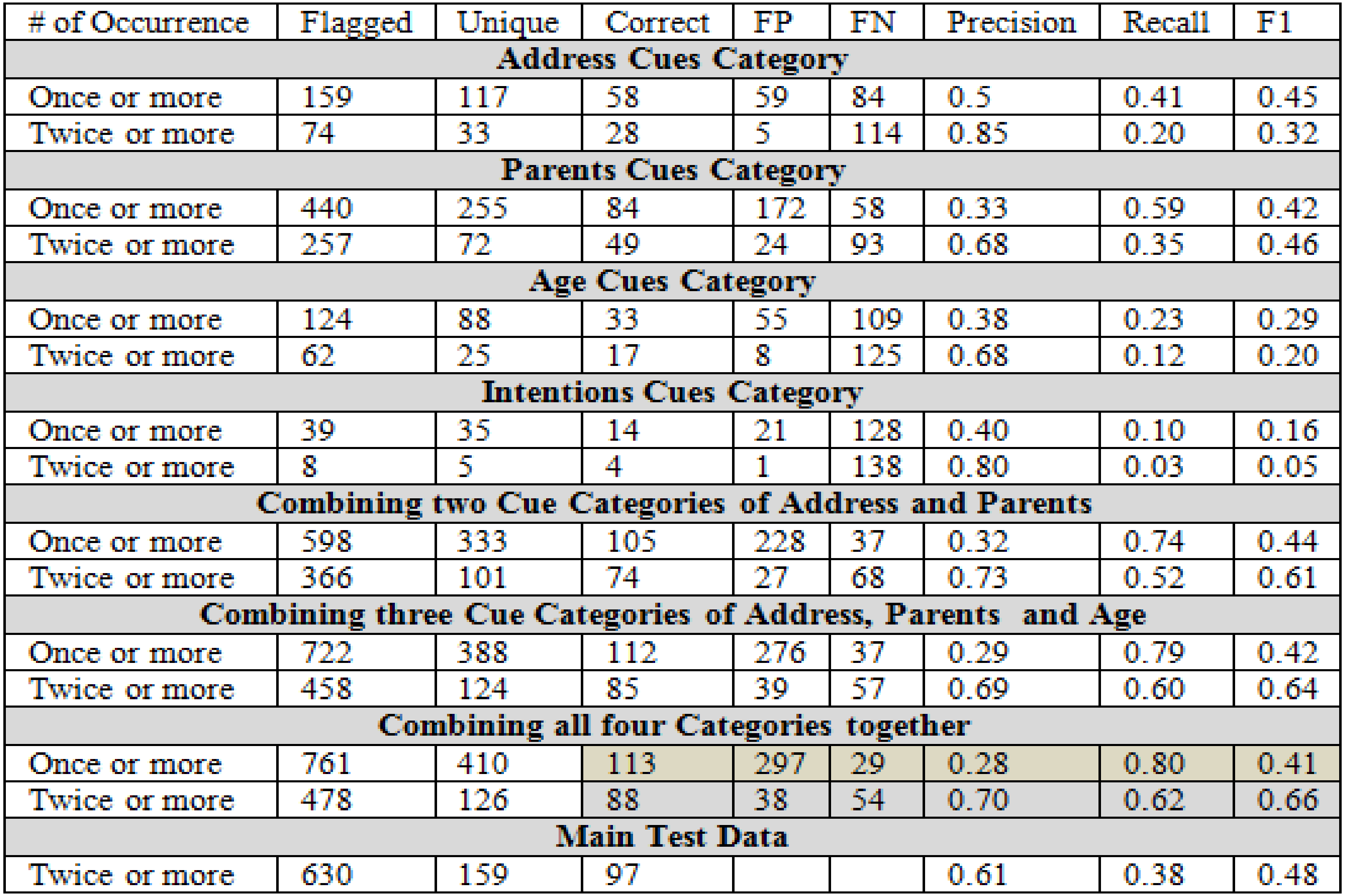
Sexual Predator Detection: Identification

- Manually extracted patterns from sample of 10 Predators’ chat



Address	Accept	13	Different spelling combination of following words: “your addres”, “ur addres”, “the addres”
	Reject	78	IT and social networking related topics such as URL, Gmail Facebook, email, e-mail, IP, Browser, …
Parents	Accept	11	Different spelling combination of following words: “your mom”, “your dad”, “your Parent”
	Reject	26	Reference to parents’ objects or characteristics such as “Ur dads car”, “Your mom’s face”, “Your mom is nice, young, etc”. IT related topics such as “Parent Class”
Age	Accept	11	Different spelling combination of following words: “you are young”, “get in trouble”, “underage”, “to jail”, “wish you were”
	Reject	33	Self-reference such as “I’m underage” Reference to the others such as sister, brother, friend Excluding, “wish you were here /with me”
Intentions	Accept	6	Different spelling combination of following words: “go down on you”, “make you come”

Sexual Predator Detection: Identification

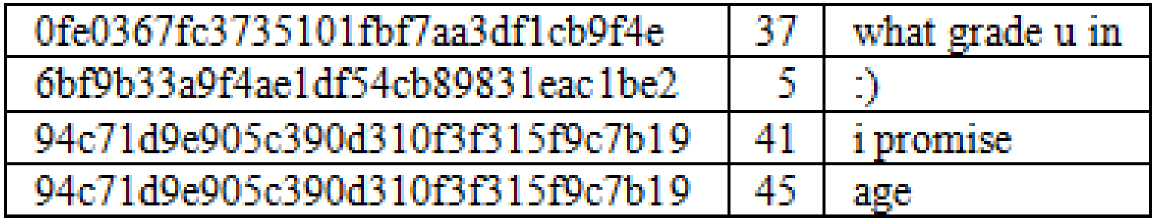


# of Occurrence	Flagged	Unique	Correct	FP	FN	Precision	Recall	F1
Address Cues Category								
Once or more	159	117	58	59	84	0.5	0.41	0.45
Twice or more	74	33	28	5	114	0.85	0.20	0.32
Parents Cues Category								
Once or more	440	255	84	172	58	0.33	0.59	0.42
Twice or more	257	72	49	24	93	0.68	0.35	0.46
Age Cues Category								
Once or more	124	88	33	55	109	0.38	0.23	0.29
Twice or more	62	25	17	8	125	0.68	0.12	0.20
Intentions Cues Category								
Once or more	39	35	14	21	128	0.40	0.10	0.16
Twice or more	8	5	4	1	138	0.80	0.03	0.05
Combining two Cue Categories of Address and Parents								
Once or more	598	333	105	228	37	0.32	0.74	0.44
Twice or more	366	101	74	27	68	0.73	0.52	0.61
Combining three Cue Categories of Address, Parents and Age								
Once or more	722	388	112	276	37	0.29	0.79	0.42
Twice or more	458	124	85	39	57	0.69	0.60	0.64
Combining all four Categories together								
Once or more	761	410	113	297	29	0.28	0.80	0.41
Twice or more	478	126	88	38	54	0.70	0.62	0.66
Main Test Data								
Twice or more	630	159	97			0.61	0.38	0.48

Sexual Predator Detection: Evaluation

- Improvements:
  - Combine all the best F1 scores from different categories
    - Parents category occurring twice or more
    - 41% to 58%
  - Populating "intentions" category

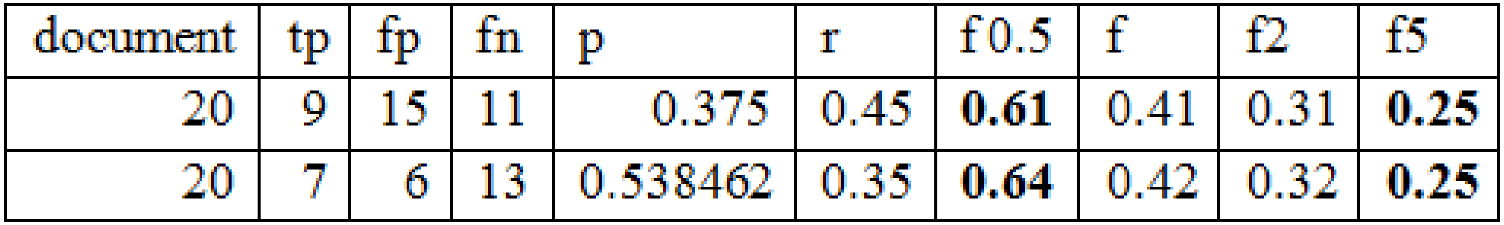
- Section two → some of these seem odd….



0fe0367fc3735101fbf7aa3df1cb9f4e	37	what grade u in
6bf9b33a9f4ae1df54cb89831eac1be2	5	:)
94c71d9e905c390d310f3f315f9c7b19	41	i promise
94c71d9e905c390d310f3f315f9c7b19	45	age

Sexual Predator Detection: Evaluation

- PAN2012: “To optimize the time of a police agent towards the "right" suspect rather than "all" the possible suspects”.
- Suppose you had 2 systems



document	tp	fp	fn	p	r	f0.5	f	f2	f5
20	9	15	11	0.375	0.45	**0.61**	0.41	0.31	**0.25**
20	7	6	13	0.538462	0.35	**0.64**	0.42	0.32	**0.25**

- Which would you prefer the police to select? (11 undetected predators, or 13?)

UNIVERSITY OF SURREY

Thank you for your attention

Anna Vartapetiance

a.vartapetiance@surrey.ac.uk

Lee Gillam

l.gillam@surrey.ac.uk

Department of Computing
University of Surrey

www.surrey.ac.uk