UNIVERSITÀ DEGLI STUDI DI PADOVA

[Towards an] Anatomy of Search Engine Performances

Nicola Ferro
@frrncl

Information Management Systems (IMS) Research Group
Department of Information Engineering (DEI)
University of Padua, Italy

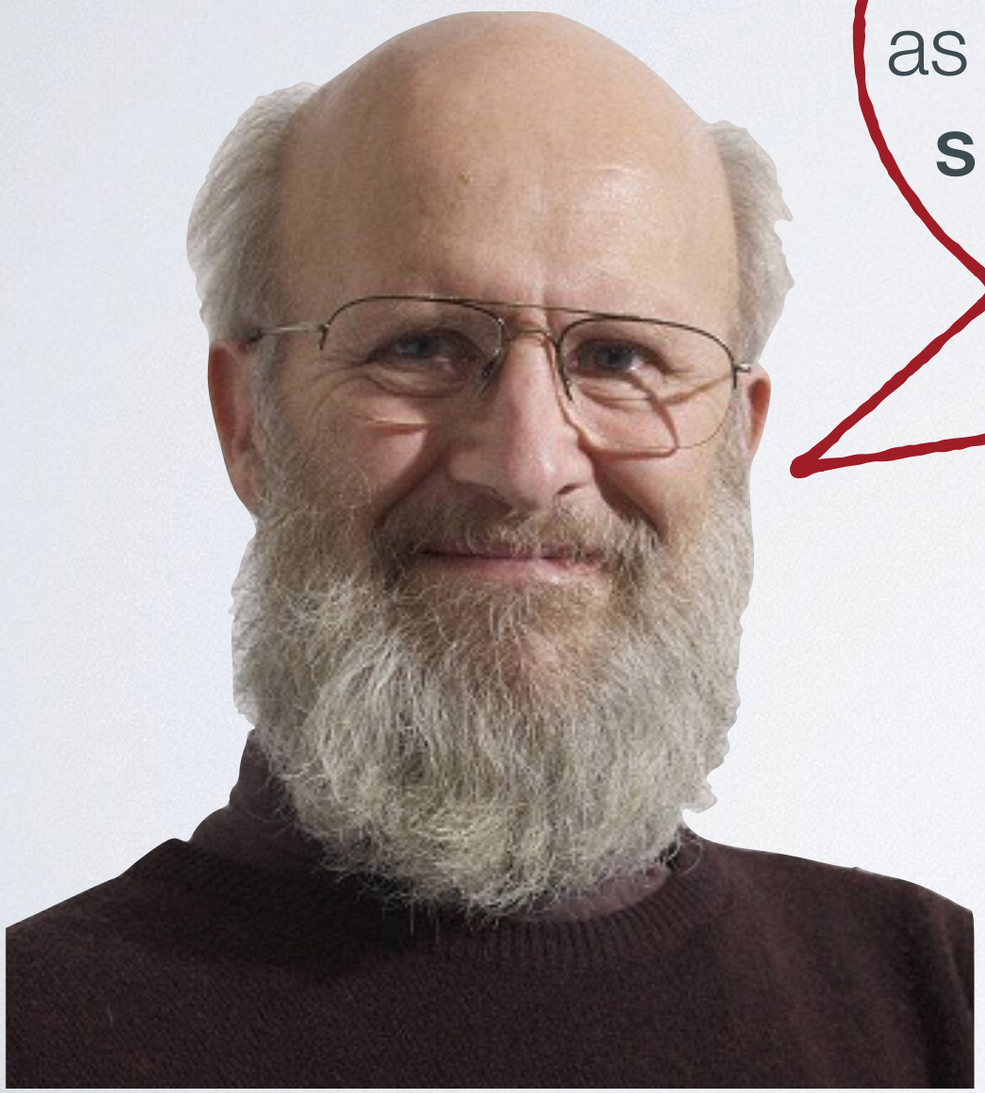
Forum for Information Retrieval Evaluation (FIRE 2016)
9 December 2016, Indian Statistical Institute (ISI), Kolkata, India

Outline

- The problem: component-based evaluation
- State of the art
- Our approach based on General Linear Mixed Models
- Experimental findings
- Conclusions and future work

The Problem

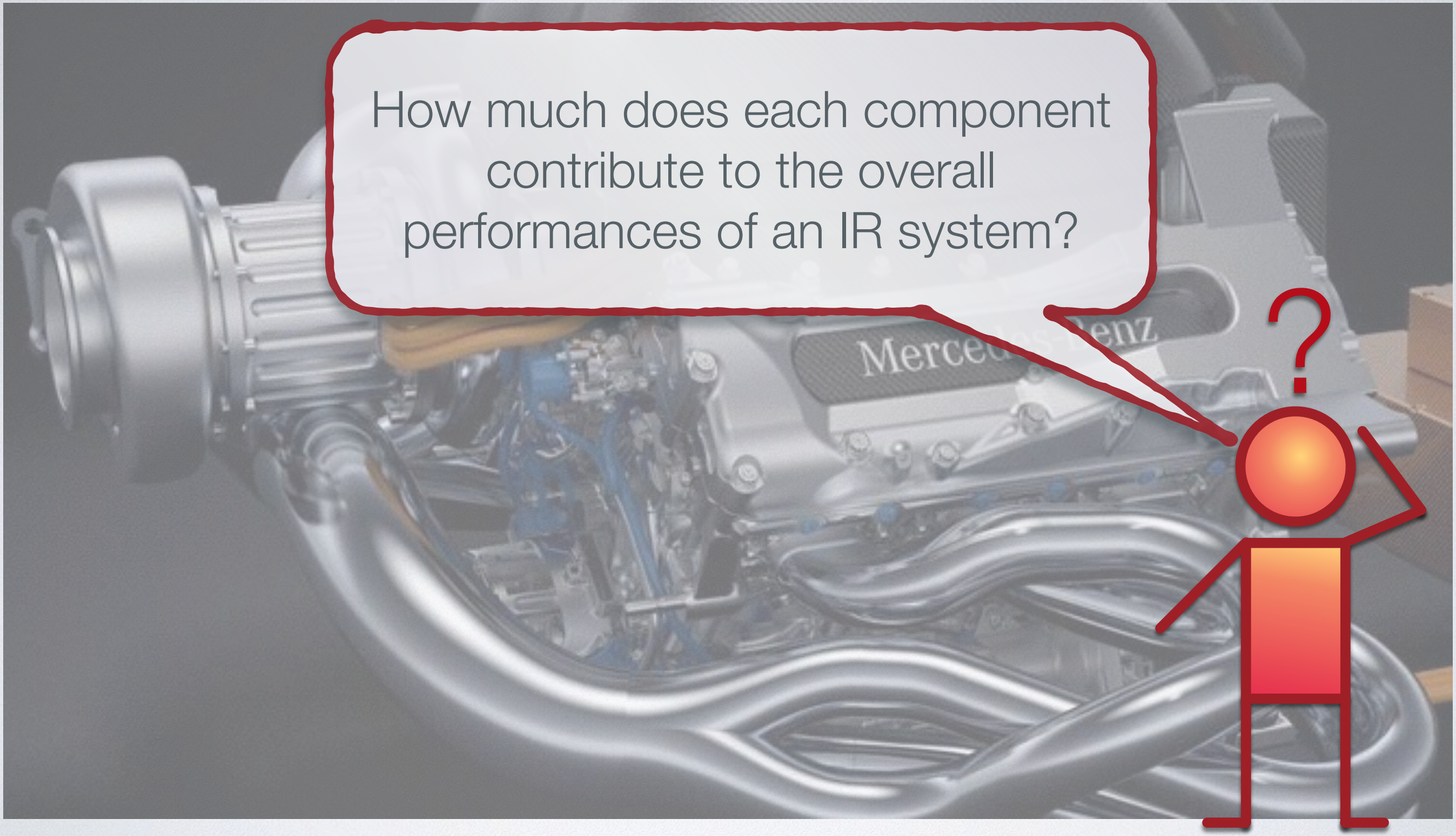
The Problem



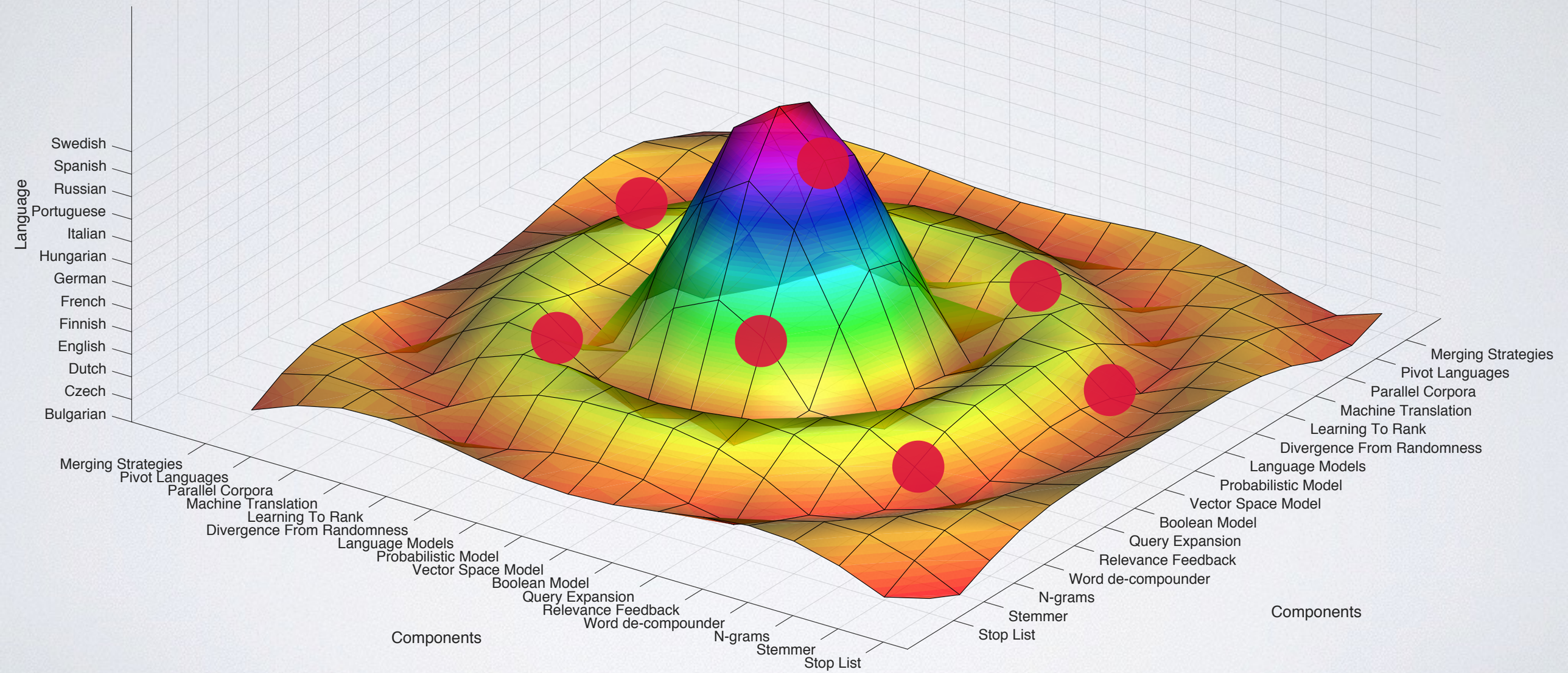
If we want to decide between **alternative** indexing **strategies**, we **must use** these strategies as part of a **complete information retrieval system**, and examine its overall performance (with each of the alternatives) directly

[Robertson, 1981]

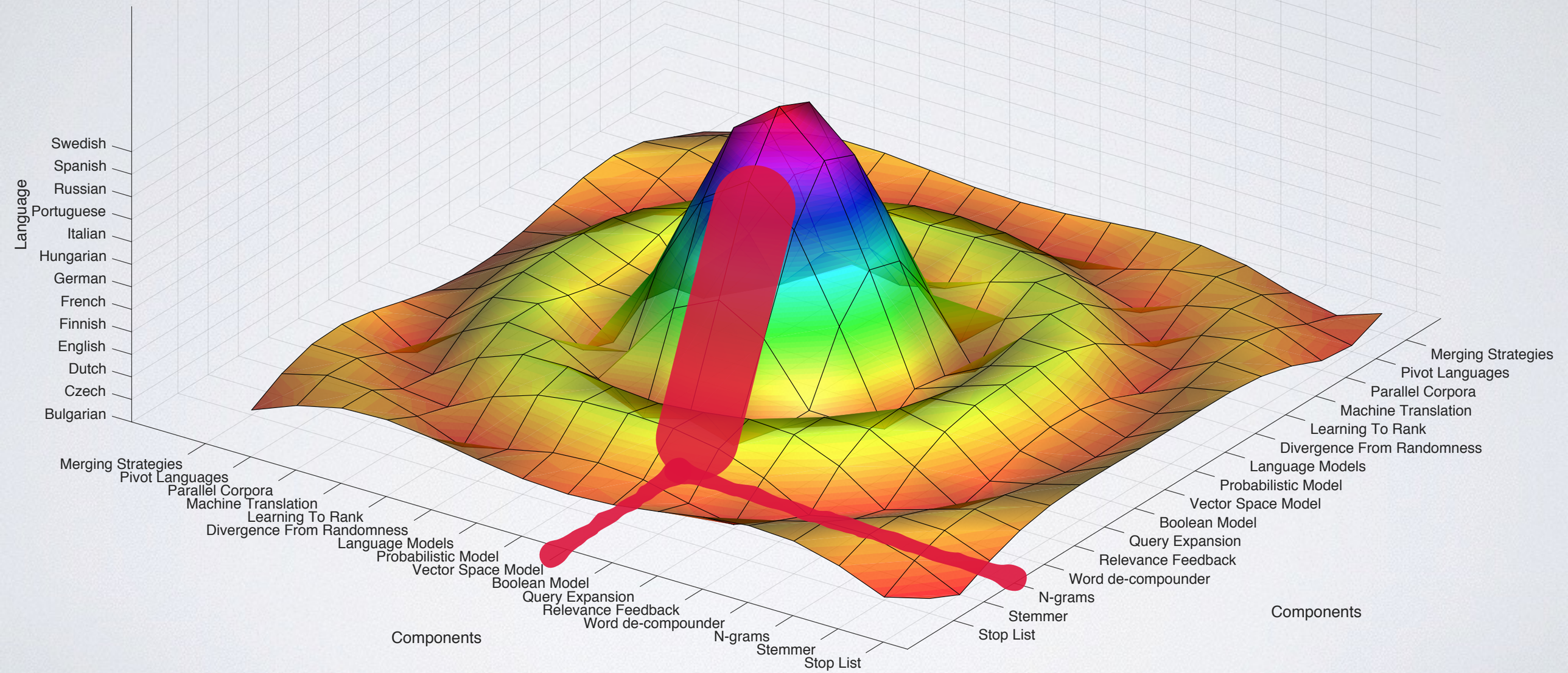
Another Side of the Problem



Typical Situation in Evaluation Campaigns



Typical Situation in Evaluation Campaigns



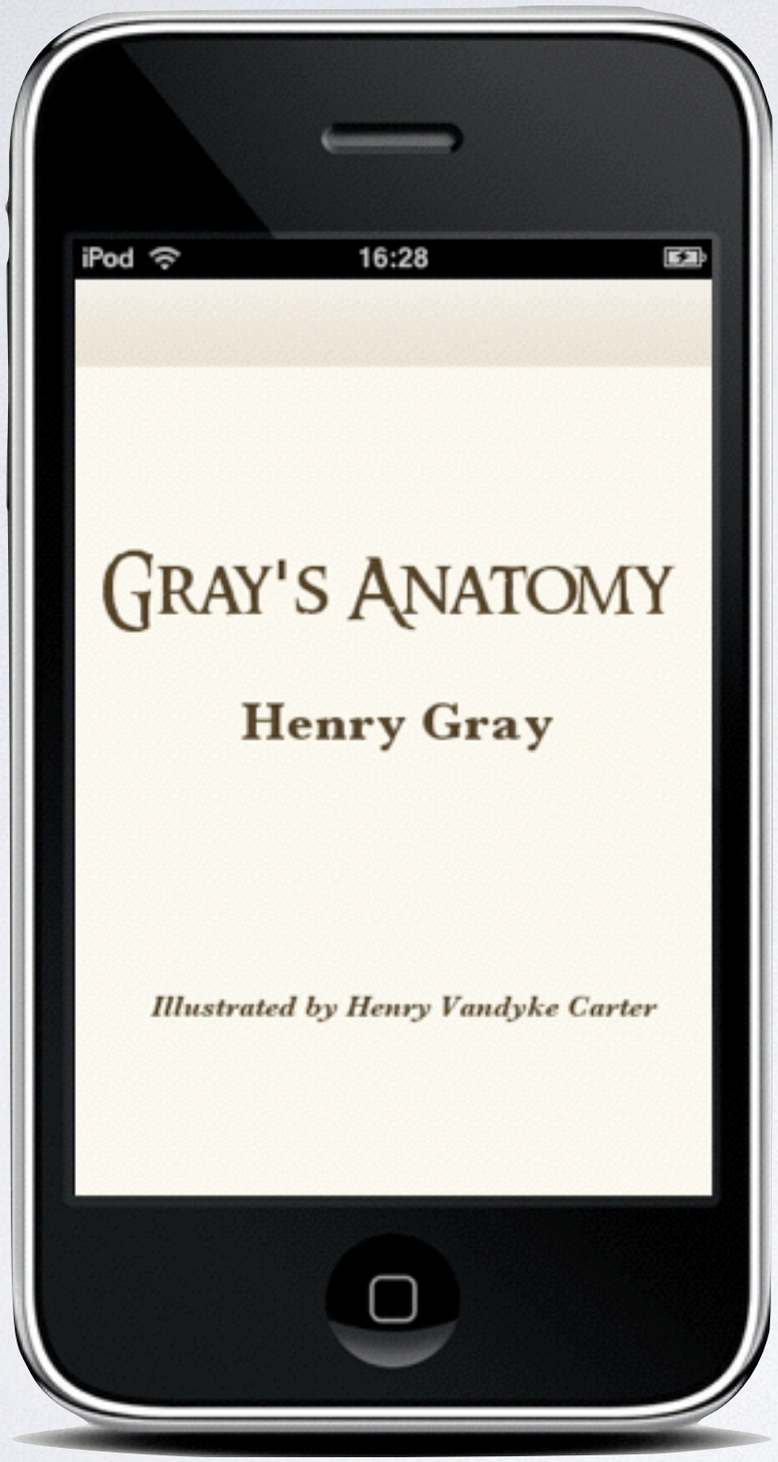
Consequences

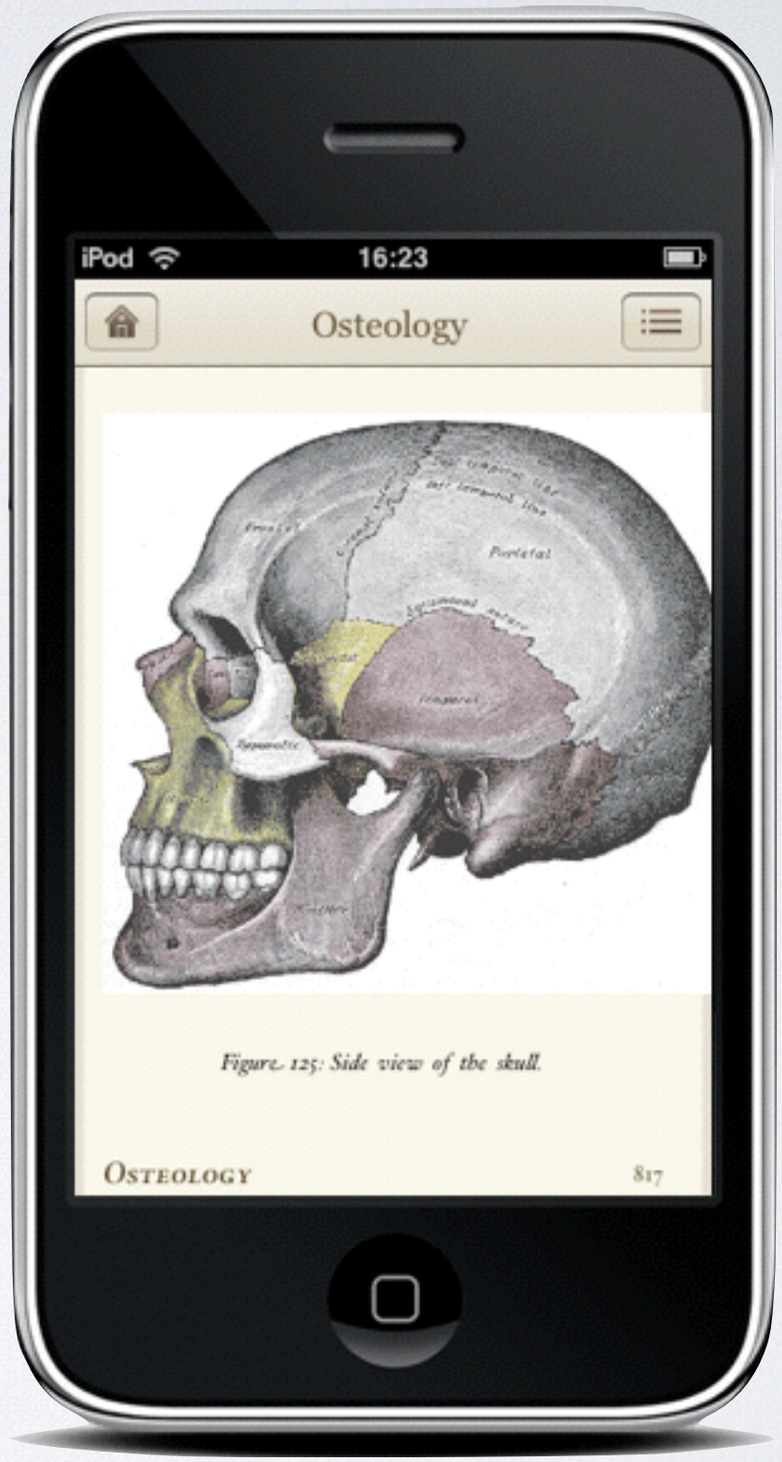
If Civil Engineering Were Like IR…

http://tinyurl.com/fuhr-clef2010

[Fuhr, 2010; Fuhr 2012]

Vision: Anatomy of IR System Performances

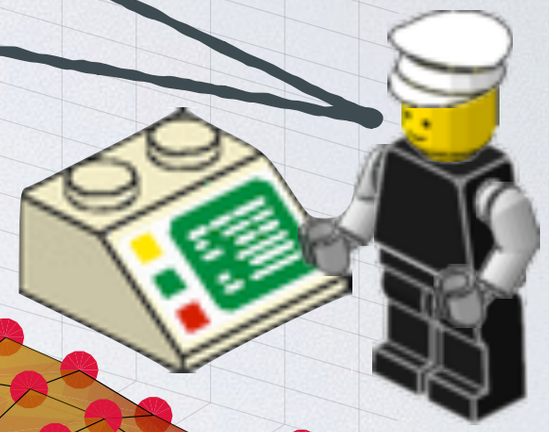




The Story so Far

Grid@CLEF: Back in 2009

To conduct a series of systematic
and comparable **grid** experiments across languages and
components by performing a **community effort** to evaluate not
only each others **components** but also their **interaction**

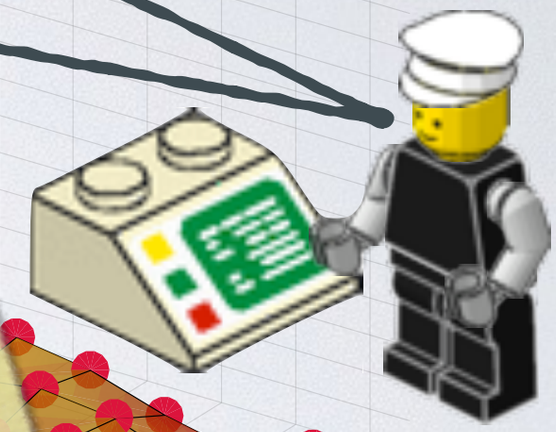


[Ferro and Harman, 2010]

Grid@CLEF: Back in 2009

To conduct a series of systematic and comparable **grid** experiments across languages and **components** by performing a **community effort** to evaluate not only each other's **components** but also their **interaction**

A proper methodology to analyze these data was **lacking**



[Ferro and Harman, 2010]

SIGIR RIGOR 2015 Open Reproducibility Challenge

- The purpose of the exercise was to invite the developers of open-source search engines to provide reproducible baselines of their systems in a common environment
- Envisaged scenarios:
  - I want to evaluate my new technique X. As a baseline, I'll use open-source search engine Y. Or alternatively, I'm building on open-source search engine Y, so I need a baseline anyway
  - How do I know what's a "reasonable" result for system Y on test collection Z? What settings should I use? (Which stopwords list? What retrieval model? What parameter settings? Etc.) How do I know if I've configured system Y correctly?
  - Correspondingly, as a reviewer of a paper that describes technique X, how do I know if the baseline is any good? Maybe the authors misconfigured system Y (inadvertently), thereby making their technique "look good" (i.e., it's a weak baseline).

[Arguello et al., 2015]

SIGIR RIGOR 2015: TREC Data

System	Model	Index	Topics			
			701–750	751–800	801–850	Combined
ATIRE	BM25	Count	0.2616	0.3106	0.2978	0.2902
ATIRE	Quantized BM25	Count + Quantized	0.2603	0.3108	0.2974	0.2897
Galago	QL	Count	0.2776	0.2937	0.2845	0.2853
Galago	SDM	Positions	0.2726	0.2911	0.3161	0.2934
Indri	QL	Positions	0.2597	0.3179	0.2830	0.2870
Indri	SDM	Positions	0.2621	0.3086	0.3165	0.2960
JASS	1B Postings	Count	0.2603	0.3109	0.2972	0.2897
JASS	2.5M Postings	Count	0.2579	0.3053	0.2959	0.2866
Lucene	BM25	Count	0.2684	0.3347	0.3050	0.3029
Lucene	BM25	Positions	0.2684	0.3347	0.3050	0.3029
MG4J	BM25	Count	0.2640	0.3336	0.2999	0.2994
MG4J	Model B	Count	0.2469	0.3207	0.3003	0.2896
MG4J	Model B+	Positions	0.2322	0.3179	0.3257	0.2923
Terrier	BM25	Count	0.2432	0.3039	0.2614	0.2697
Terrier	DPH	Count	0.2768	0.3311	0.2899	0.2994
Terrier	DPH + Bo1 QE	Count (inc direct)	0.3037	0.3742	0.3480	0.3422
Terrier	DPH + Prox SD	Positions	0.2750	0.3297	0.2897	0.2983

SIGIR RIGOR 2015: CLEF Data

System	Model	Stop	Stem	bg	de	es	fa	fi	fr
Terrier	BM25			0.2092	0.2733	0.3627	0.4033	0.3464	–
Terrier	BM25	✓		0.2081	0.2742	0.3656	0.4022	0.3392	–
Terrier	BM25		✓	–	0.3194	0.4347	–	0.4339	–
Terrier	BM25	✓	✓	–	0.3215	0.4356	–	0.4278	–
Terrier	Hiemstra LM			0.1647	0.2520	0.3016	0.3140	0.3125	–
Terrier	Hiemstra LM	✓		0.1640	0.2561	0.3081	0.3193	0.3156	–
Terrier	Hiemstra LM		✓	–	0.2753	0.3673	–	0.3639	–
Terrier	Hiemstra LM	✓	✓	–	0.2801	0.3783	–	0.3636	–
Terrier	PL2			0.2043	0.2625	0.3486	0.4081	0.3316	–
Terrier	PL2	✓		0.2009	0.2658	0.3572	0.4061	0.3388	–
Terrier	PL2		✓	–	0.3080	0.4168	–	0.4222	–
Terrier	PL2	✓	✓	–	0.3102	0.4211	–	0.4152	–
Terrier	TFIDF			0.2071	0.2709	0.3597	0.4050	0.3457	–
Terrier	TFIDF	✓		0.2083	0.2723	0.3658	0.4053	0.3393	–
Terrier	TFIDF		✓	–	0.3185	0.4313	–	0.4354	–
Terrier	TFIDF	✓	✓	–	0.3167	0.4355	–	0.4269	–
Lucene	BM25	✓	✓	–	0.3126	0.4251	0.4158	–	0.3865
Indri	LM Dirichlet	✓	✓	0.2051	0.1365	0.3334	0.3735	–	0.1444

System	Model	Stop	Stem	hu	it	nl	pt	ru	sv
Terrier	BM25			0.2115	0.3233	0.3958	0.3250	0.3666	0.3384
Terrier	BM25	✓		0.2178	0.3182	0.3974	0.3255	0.3449	0.3371
Terrier	BM25		✓	0.3175	0.3619	0.4209	0.3250	0.4740	0.3817
Terrier	BM25	✓	✓	0.3254	0.3591	0.4234	0.3255	0.4753	0.3886
Terrier	Hiemstra LM			0.1642	0.2778	0.3454	0.2738	0.2922	0.3113
Terrier	Hiemstra LM	✓		0.1685	0.2820	0.3523	0.2742	0.2949	0.3160
Terrier	Hiemstra LM		✓	0.2559	0.3061	0.3585	0.2738	0.3891	0.3372
Terrier	Hiemstra LM	✓	✓	0.2656	0.3092	0.3680	0.2742	0.3960	0.3402
Terrier	PL2			0.2060	0.3110	0.3792	0.3183	0.3433	0.3149
Terrier	PL2	✓		0.2091	0.3090	0.3832	0.3184	0.3288	0.3222
Terrier	PL2		✓	0.3040	0.3521	0.4042	0.3183	0.4737	0.3604
Terrier	PL2	✓	✓	0.3179	0.3472	0.4088	0.3184	0.4711	0.3708
Terrier	TFIDF			0.2107	0.3238	0.3946	0.3230	0.3643	0.3344
Terrier	TFIDF	✓		0.2181	0.3205	0.3975	0.3258	0.3403	0.3354
Terrier	TFIDF		✓	0.3105	0.3675	0.4222	0.3230	0.4764	0.3789
Terrier	TFIDF	✓	✓	0.3252	0.3649	0.4253	0.3258	0.4647	0.3869
Lucene	BM25	✓	✓	0.3233	0.3486	0.4172	–	0.4717	0.3775
Indri	LM Dirichlet	✓	✓	0.2381	0.0984	0.2486	–	0.2991	0.3265

SIGIR RIGOR 2015: CLEF Data

System	Model	Stop	Stem	bg	de	es	fa	fi	fr
Terrier	BM25			0.2092	0.2733	0.3627	0.4033	0.3464	–
Terrier	BM25	✓		0.2081	0.2742	0.3656	0.4022	0.3392	–
Terrier	BM25		✓	–	0.3194	0.4347	–	0.4339	–
Terrier	BM25	✓	✓	–	0.3215	0.4356	–	0.4278	–
Terrier	Hiemstra LM			0.1647	0.2520	0.3016	0.3140	0.3125	–
Terrier	Hiemstra LM	✓		0.1640	0.2561	0.3081	0.3193	0.3156	–
Terrier	Hiemstra LM		✓	–	0.2753	0.3673	–	0.3639	–
Terrier	Hiemstra LM	✓	✓	–	0.2801	0.3783	–	0.3636	–
Terrier	PL2			0.2043	0.2625	0.3486	0.4081	0.3316	–
Terrier	PL2	✓		0.2009	0.2658	0.3572	0.4061	0.3388	–
Terrier	PL2		✓	–	0.3080	0.4168	–	0.4222	–
Terrier	PL2	✓	✓	–	0.3102	0.4211	–	0.4152	–
Terrier	TFIDF			0.2071	0.2709	0.3597	0.4050	0.3457	–
Terrier	TFIDF	✓		0.2083	0.2733	0.3658	0.4053	0.3393	–
Terrier	TFIDF		✓	–	0.3185	0.4313	–	0.4354	–
Terrier	TFIDF	✓	✓	–	0.3167	0.4355	–	0.4269	–
Lucene	BM25		✓	0.3126	0.4258	0.4158	–		0.3865
Indri	LM Dirichlet		✓	0.2051	0.1365	0.3333	0.3735	–	0.1444

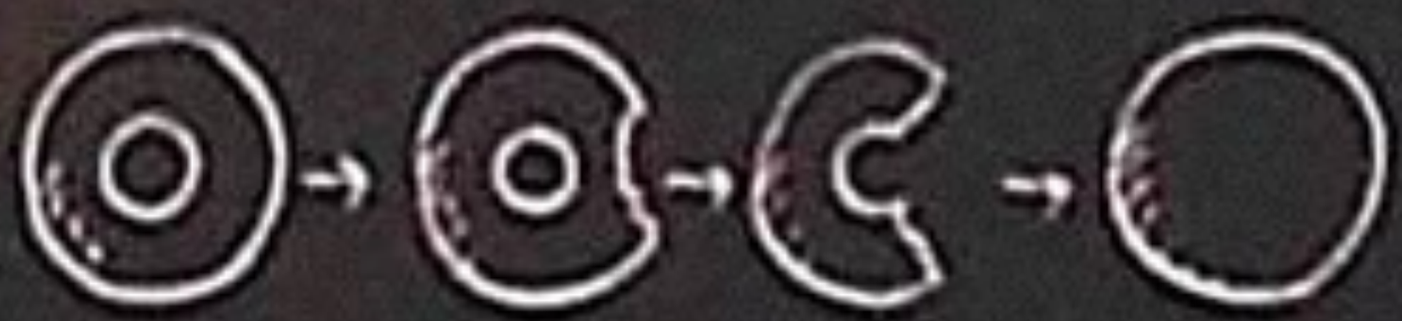
System	Model	Stop	Stem	it	nl	pt	ru	sv
Terrier	BM25			0.2115	0.3295	0.3250	0.3666	0.3384
Terrier	BM25	✓		0.2111	0.3274	0.3974	0.3255	0.3449	0.3371
Terrier	BM25		✓	0.3256	0.3619	0.4209	0.4740	0.3817
Terrier	BM25	✓	✓	0.3254	0.3591	0.4205	0.4753	0.3886
Terrier	Hiemstra LM			0.1642	0.2738	0.2922	0.3113
Terrier	Hiemstra LM	✓		0.1685	0.3523	0.2742	0.2949	0.3160
Terrier	Hiemstra LM		✓	0.2659	0.3061	0.3585	0.2738	0.3891	0.3372
Terrier	Hiemstra LM	✓	✓	0.2656	0.3092	0.3680	0.2742	0.3960	0.3402
Terrier	PL2			0.2060	0.3110	0.3792	0.3183	0.3433	0.3149
Terrier	PL2	✓		0.2091	0.3090	0.3832	0.3184	0.3288	0.3222
Terrier	PL2		✓	0.3040	0.3521	0.4042	0.3183	0.4737	0.3604
Terrier	PL2	✓	✓	0.3179	0.3472	0.4088	0.3184	0.4711	0.3708
Terrier	TFIDF			0.2107	0.3238	0.3946	0.3230	0.3643	0.3344
Terrier	TFIDF	✓		0.2181	0.3205	0.3975	0.3258	0.3403	0.3354
Terrier	TFIDF		✓	0.3105	0.3675	0.4222	0.3230	0.4764	0.3789
Terrier	TFIDF	✓	✓	0.3252	0.3649	0.4253	0.3258	0.4647	0.3869
Lucene	BM25	✓	✓	0.3233	0.3486	0.4172	–	0.4717	0.3775
Indri	LM Dirichlet	✓	✓	0.2381	0.0984	0.2486	–	0.2991	0.3265

A proper methodology to analyze these data is still **lacking**

$$M(H^{\circ}) = \pi \left(\frac{1}{137}\right)^{8} \sqrt{\frac{hc}{G}}$$

$$3987^{12} + 4365^{12} = 4472^{12}$$

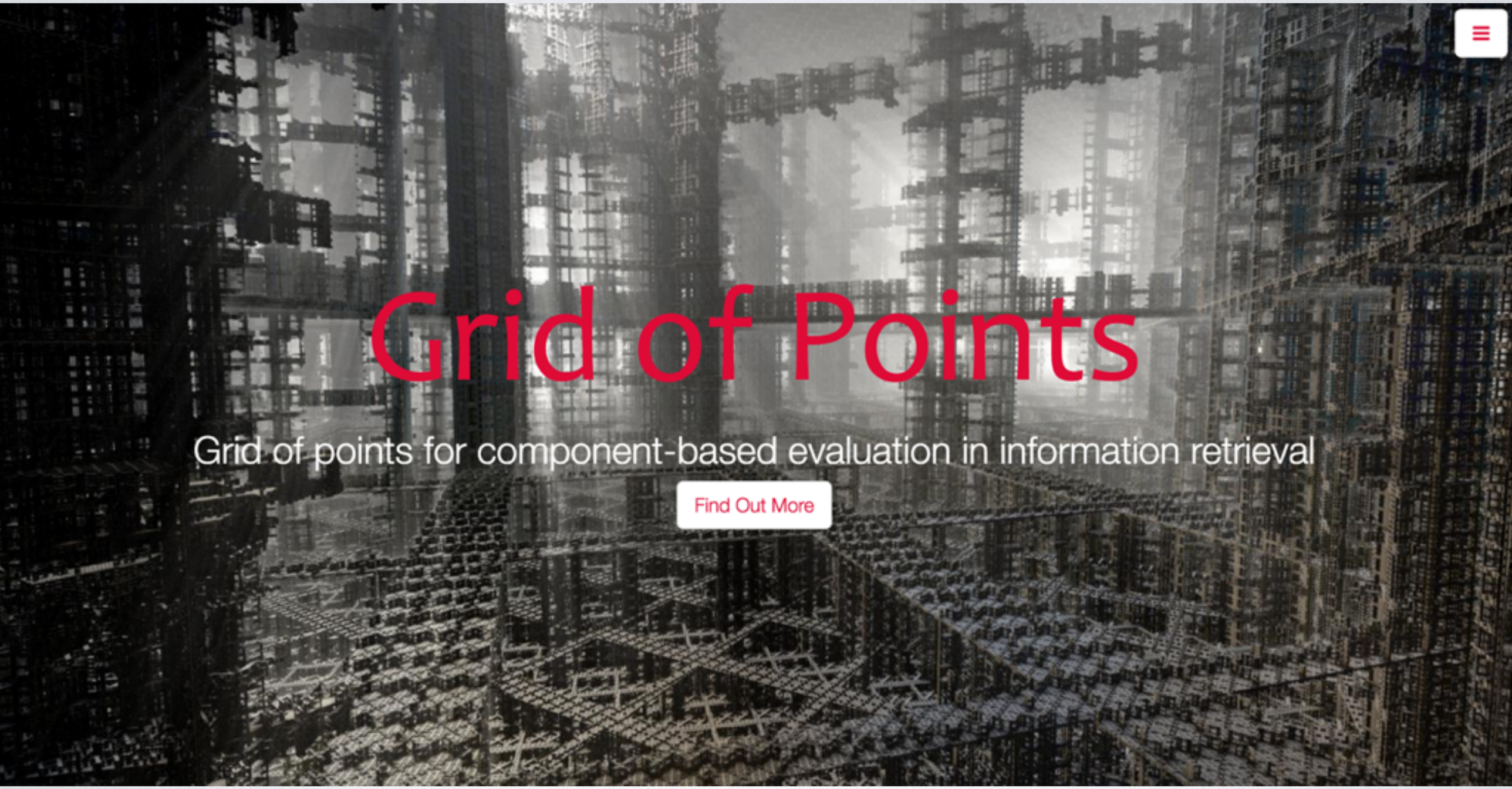
$$\Omega(t_{\cdot}) > 1$$



Our Approach

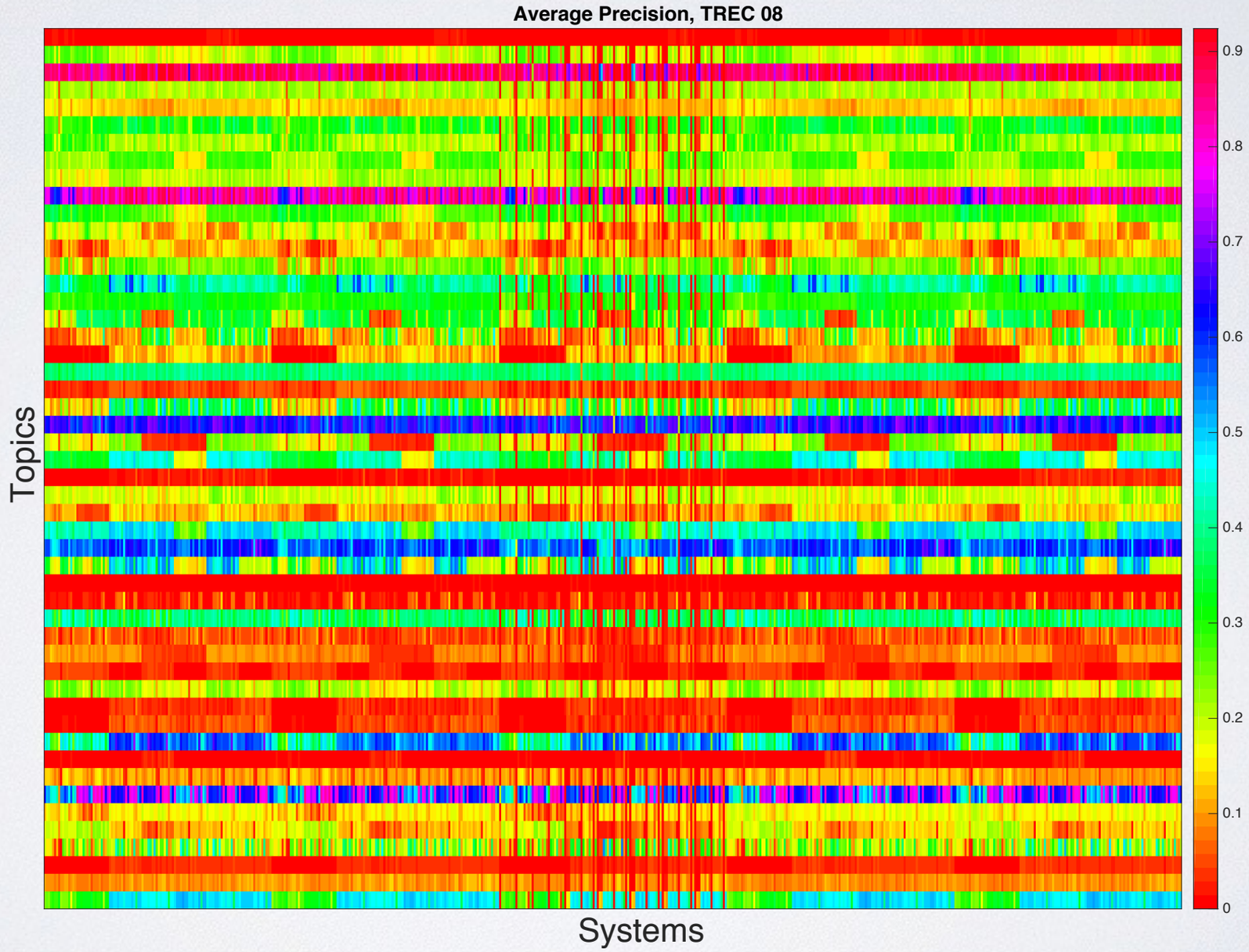
Grid of Points (GoP)

http://gridofpoints.dei.unipd.it/



What Does Affect Performances?

$$\text{system performances} = \text{topic effect} + \text{system effect} + \sout{\text{topic/system interaction effect}}$$



General Linear Mixed Models

A General Linear Mixed Model (GLMM) explains the variation of a dependent variable Y ("Data") in terms of a controlled variation of independent variables ("Model") in addition to a residual uncontrolled variation ("Error")

Data = Model + Error

- General: both continuous and categorical variables
- Linear: the model is expressed as a linear combination of factors
- Mixed: both fixed and random factors
- Experiment design
  - independent vs repeated measures
  - factorial vs nested

Single Factor Repeated Measures Design

$$Y_{ij} = \underbrace{\mu.. + \tau_i + \alpha_j}_{\text{Model}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}$$

Factor A (Systems)

Subjects (Topics)	A_1	A_2	$\cdots$	A_p	
T_1'	Y_{11}	Y_{12}	$\cdots$	Y_{1p}	$\mu_{1.}$
T_2'	Y_{21}	Y_{22}	$\cdots$	Y_{2p}	$\mu_{2.}$
$\vdots$	$\vdots$	$\vdots$	Y_{ij}	$\vdots$	$\mu_{i.}$
T_n'	Y_{n1}	Y_{n2}	$\cdots$	Y_{np}	$\mu_{n.}$
	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.j}$	$\mu_{.p}$	$\mu..$

Assessment: Strength of Association

The **Effect-size Measure** or **Strength of Association (SOA)** is a standardized index, independent of sample size, which quantifies the relationship between explanatory and response variables

$$\hat{\omega}^2_{\langle fact \rangle} = \frac{df_{fact}(F_{fact} - 1)}{df_{fact}(F_{fact} - 1) + pn}$$

Rule of thumb

- **Large effect**: 0.14 and above
- **Medium effect**: 0.06–0.14
- **Small effect**: 0.01–0.06

Assessment: Type I and Type II Errors

- **Type I Error**: occurs when a true null hypothesis is rejected and the significance level α is the probability of committing a Type I error
  - A Type I error identifies a false effect that can misdirect theory development and empirical effort, and requires empirical and/or theoretical effort to remedy
- **Type II Error**: occurs when a false null hypothesis is accepted and it is concerned with the capability of the conducted experiment to actually detect the effect under examination
  - Type II errors are often overlooked because if they occur, although a real effect is missed, no misdirection occurs and further experimentation is very likely to reveal the effect

Assessment: Power

The **power** is the probability of correctly rejecting a false null hypothesis when an experimental hypothesis is true

$$\text{Power} = 1 - \beta$$

where β (typically $\beta = 0.2$) is the Type II error rate.

- Compute the effect size parameter

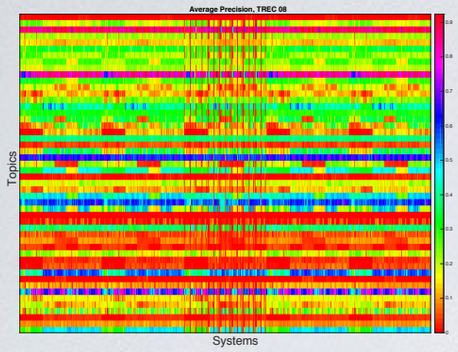
$$\phi = \sqrt{n \cdot \frac{\hat{\omega}^2_{\langle fact \rangle}}{1 - \hat{\omega}^2_{\langle fact \rangle}}}$$

- Compare it with its tabulated values for a given Type I error rate α to determine β
  - We used G*Power (**`http://www.gpower.hhu.de/`**)

Experimental Setup

- A Grid of Points (GoP) consisting of 560 system has been created with all the possible combinations of the following components
  - **Stop list**: nostop, indri, lucene, smart, terrier;
  - **Lexical Unit Generator (LUG)**:
    - nolug, weak Porter, Porter, Krovetz, Lovins;
    - nolug, 4grams, 5grams;
  - **Model**: BB2, BM25, DFRBM25, DFRee, DLH, DLH13, DPH, HiemstraLM, IFB2, InL2, InexpB2, InexpC2, LGD, LemurTFIDF, PL2, TFIDF.
- Experimental collections: TREC 5, 6, 7, and 8 Adhoc
- Measures: AP, P@10, RBP, nDCG@20, ERR@20

Single Factor, TREC 08, AP



Source	SS	DF	MS	F	p-value
Topics′	820.99	49	16.75	694.7235	0
Systems	36.44	399	0.09	7.4464	0
Error	88.20	19551	0.0045		
Total	945.63	19999			

- Topics explain a large portion of the total variance
  - consistent with previous findings [Tague-Sutcliffe and Blustein, 1994]
- The effect of the IR systems is statistically significant
- The sum of squares of the error is not negligible since it contains both the variance of the unexplained topics/systems interaction effect and the the other uncontrolled sources of variance
- The power of the experiment is 1 with a Type I error probability $\alpha = 0.05$ indicating that we are observing effects in a reliable way

Single Factor: Strength of Association

Collection	LUG	Effects	AP	P@10	RBP	nDCG@20	ERR@20
TREC 05	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	0.1223 (0.00)	**0.2023** (0.00)	**0.1970** (0.00)	**0.1879** (0.00)	**0.1406** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	0.0794 (0.00)	0.1178 (0.00)	0.1349 (0.00)	0.1200 (0.00)	0.1063 (0.00)
TREC 06	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.2108** (0.00)	**0.2458** (0.00)	**0.2716** (0.00)	**0.2742** (0.00)	**0.2377** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	0.1350 (0.00)	**0.1496** (0.00)	**0.1597** (0.00)	**0.1725** (0.00)	**0.1469** (0.00)
TREC 07	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.2155** (0.00)	**0.2568** (0.00)	**0.2894** (0.00)	**0.2977** (0.00)	**0.2445** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.1502** (0.00)	**0.1658** (0.00)	**0.1920** (0.00)	**0.1898** (0.00)	**0.1480** (0.00)
TREC 08	Stemmers	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.2774** (0.00)	**0.2780** (0.00)	**0.3025** (0.00)	**0.3118** (0.00)	**0.2484** (0.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Systems} \rangle}$	**0.1758** (0.00)	**0.1907** (0.00)	**0.2006** (0.00)	**0.2135** (0.00)	**0.1530** (0.00)

- Despite the high variance of the topics, the system effect sizes are generally large and significant. This is consistent across all the collections and measures
- System effect sizes of stemmer runs group systems are large (> 0.14) for all the collections and measures with the solely exception of AP for TREC 05
- For the n-grams runs group we can see that the system effect sizes are consistently smaller than those of the stemmer group
  - this, supports the observation that “for English, n-grams indexing has no strong impact” [Büttcher et al, 2010]
- System effect sizes are higher when nDCG@20 is used, followed by RBP, P@10, AP and ERR@20

Single Factor: Discriminative Power

Group		TREC 05	TREC 06	TREC 07	TREC 08
stemmer	AP	3011	.2748	.3591	.4743
	P@10	.3774	.2687	.3222	.3171
	RBP	.3152	.2589	.3302	.3422
	nDCG@20	.3448	.2698	.3169	.3834
	ERR@20	.2014	.2235	.2096	.2388
n-grams	AP	.3180	.3553	.5184	.3498
	P@10	.3025	.2656	.3660	.2977
	RBP	.3852	.2539	.4193	.2797
	nDCG@20	.3260	.3130	.4292	.2938
	ERR@20	.2832	.1978	.2549	.2416

- Hypothesis on measures impact on SoA
  - **Discriminative power**: if a measure is less discriminative than another one, it could be able to grasp less variance in the system effect
  - **User model**: different user models mean looking at (very) different angles of system performances and this can change the explained variance
- We can see that there is some agreement between the system effect sizes for a measure and its discriminative power
  - ERR@20 explains less system variance than the other measures and this can be explained by its discriminative power which is the lowest amongst all measures
  - RBP and nDCG@20 have both comparable discriminative power and close system effect sizes
- The main exception is AP which typically has the highest discriminative power but the smallest system effect size
  - this could be due to the user model behind AP, which is quite different from the one of the other measures and may counterbalance the higher discriminative power leading to a final lower system effect size

Three Factors Repeated Measures Design

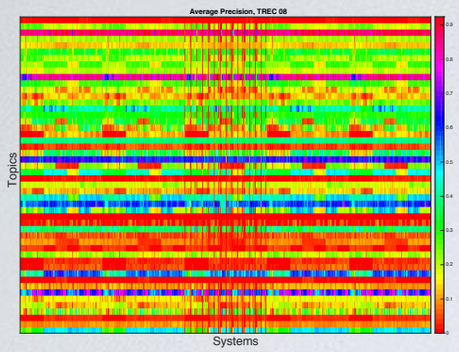
$$Y_{ijkl} = \underbrace{\mu.... + \tau_i + \alpha_j + \beta_k + \gamma_l}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \beta\gamma_{kl} + \alpha\beta\gamma_{jkl}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijkl}}_{\text{Error}}$$

Factor A (Stop Lists)
Factor B (Lexical Unit Generator)

Factor C (Models)
Subjects (Topics)

		A_1				A_2				⋯				A_p			
		B_1	B_2	⋯	B_q	B_1	B_2	⋯	B_q					B_1	B_2	⋯	B_q
C_1	T_1'	Y_{1111}	Y_{1121}		Y_{11q1}	Y_{1211}	Y_{1221}		Y_{12q1}					Y_{1p11}	Y_{1p21}		Y_{1pq1}
	T_2'	Y_{2111}	Y_{2121}	⋯	Y_{21q1}	Y_{2211}	Y_{2221}	⋯	Y_{22q1}	⋯				Y_{2p11}	Y_{2p21}	⋯	Y_{2pq1}
		⋮	⋮		⋮	⋮	⋮		⋮					⋮	⋮		⋮
	T_n'	Y_{n111}	Y_{n121}		Y_{n1q1}	Y_{n211}	Y_{n221}		Y_{n2q1}					Y_{np11}	Y_{np21}		Y_{npq1}
C_2	T_1'	Y_{1112}	Y_{1122}		Y_{11q2}	Y_{1212}	Y_{1222}		Y_{12q2}					Y_{1p12}	Y_{1p22}		Y_{1pq2}
	T_2'	Y_{2112}	Y_{2122}	⋯	Y_{21q2}	Y_{2212}	Y_{2222}	⋯	Y_{22q2}	⋯				Y_{2p12}	Y_{2p22}	⋯	Y_{2pq2}
		⋮	⋮		⋮	⋮	⋮		⋮					⋮	⋮		⋮
	T_n'	Y_{n112}	Y_{n122}		Y_{n1q2}	Y_{n212}	Y_{n222}		Y_{n2q2}					Y_{np12}	Y_{np22}		Y_{npq2}
⋮	T_1'	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋱				⋮	⋮	⋱	⋮
	T_2'																
	T_n'																
C_r	T_1'	Y_{111r}	Y_{112r}		Y_{11qr}	Y_{121r}	Y_{122r}		Y_{12qr}					Y_{1p1r}	Y_{1p2r}		Y_{1pqr}
	T_2'	Y_{211r}	Y_{212r}	⋯	Y_{21qr}	Y_{221r}	Y_{222r}	⋯	Y_{22qr}	⋯				Y_{2p1r}	Y_{2p2r}	⋯	Y_{2pqr}
		⋮	⋮		⋮	⋮	⋮		⋮					⋮	⋮		⋮
	T_n'	Y_{n11r}	Y_{n12r}		Y_{n1qr}	Y_{n21r}	Y_{n22r}		Y_{n2qr}					Y_{np1r}	Y_{np2r}		Y_{npqr}

Three Factors, TREC 08, AP



Source	SS	DF	MS	F	p
Topics′	820.99	49	16.75	3713.90	0.00
Stop list	9.89	4	2.47	548.06	0.00
Stemmer	4.16	4	1.04	230.76	0.00
Model	5.16	15	0.3443	76.32	0.00
Stop list*Stemmer	0.05	16	0.03	0.67	0.83
Stop list*Model	17.01	60	0.28	62.84	0.00
Stemmer*Model	0.07	60	0.001	0.26	1.00
Stop list*Stemmer*Model	0.09	240	0.00	0.08	1.00
Error	88.20	19551	0.005		
Total	945.63	19999			

- First order effects are all significant
- The stop list explains more variance than the model and the stemmer is the component with the lowest impact
- Only the stoplist*model effect is significant explaining a tangible portion of the systems variance
- The power for the main effects is 0.97 for the stop list, **0.66** for the stemmer and 0.99 for the model with a Type I error probability $\alpha = 0.05$

Three Factors: Strength of Association

Collection	LUG	Effects	AP	P@10	RBP	nDCG@20	ERR@20
TREC 08	Stemmers	$\hat{\omega}_{\langle \text{Stop Lists} \rangle}$	**0.0986** (0.00)	**0.0913** (0.00)	**0.1000** (0.00)	**0.1006** (0.00)	**0.0799** (0.00)
		$\hat{\omega}^2_{\langle \text{Stemmers} \rangle}$	0.0439 (0.00)	0.0165 (0.00)	0.0190 (0.00)	0.0268 (0.00)	0.0071 (0.00)
		$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0535 (0.00)	**0.0615** (0.00)	**0.0666** (0.00)	**0.0707** (0.00)	0.0521 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0003 (0.83)	-0.0005 (0.98)	-0.0005 (0.98)	-0.0006 (0.99)	-0.0004 (0.95)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.1565** (0.00)	**0.1765** (0.00)	**0.1969** (0.00)	**0.2006** (0.00)	**0.1622** (0.00)
		$\hat{\omega}^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0022 (1.00)	-0.0014 (0.99)	-0.0020 (1.00)	-0.0018 (1.00)	-0.0016 (0.99)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0111 (1.00)	-0.0105 (1.00)	-0.0110 (1.00)	-0.0110 (1.00)	-0.0102 (1.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	0.0396 (0.00)	0.0423 (0.00)	0.0445 (0.00)	0.0479 (0.00)	0.0304 (0.00)
		$\hat{\omega}^2_{\langle n\text{-grams} \rangle}$	0.0037 (0.00)	0.0031 (0.00)	0.0008 (0.00)	0.0023 (0.00)	0.0093 (0.00)
		$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0550 (0.00)	0.0545 (0.00)	0.0548 (0.00)	**0.0637** (0.00)	0.0307 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times n\text{-grams} \rangle}$	0.0035 (0.00)	0.0023 (0.00)	0.0024 (0.00)	0.0029 (0.00)	0.0032 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.0928** (0.00)	**0.1129** (0.00)	**0.1231** (0.00)	**0.1277** (0.00)	**0.0940** (0.00)
		$\hat{\omega}^2_{\langle n\text{-grams} \times \text{IR Models} \rangle}$	0.0080 (0.00)	0.0050 (0.00)	0.0059 (0.00)	0.0050 (0.00)	0.0040 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times n\text{-grams} \times \text{IR Models} \rangle}$	-0.0038 (0.99)	-0.0040 (0.99)	-0.0032 (0.99)	-0.0034 (0.99)	-0.0028 (0.99)

- For the stemmer group, the stop list has always a higher SoA than the IR model and the stemmer and the stop list have a medium effect size
  - The stemmer*model interaction effect which is never significant
- N-grams tend to reduce the stop list effect and to increase the IR model one
  - The n-grams*model interaction effect which is small but statistically significant
- The stop list*model interaction effect is always the biggest of effects (main and interaction ones)
- Not all the measures detect similarly well components effects, e.g. ERR@20 almost ignores the stemmer

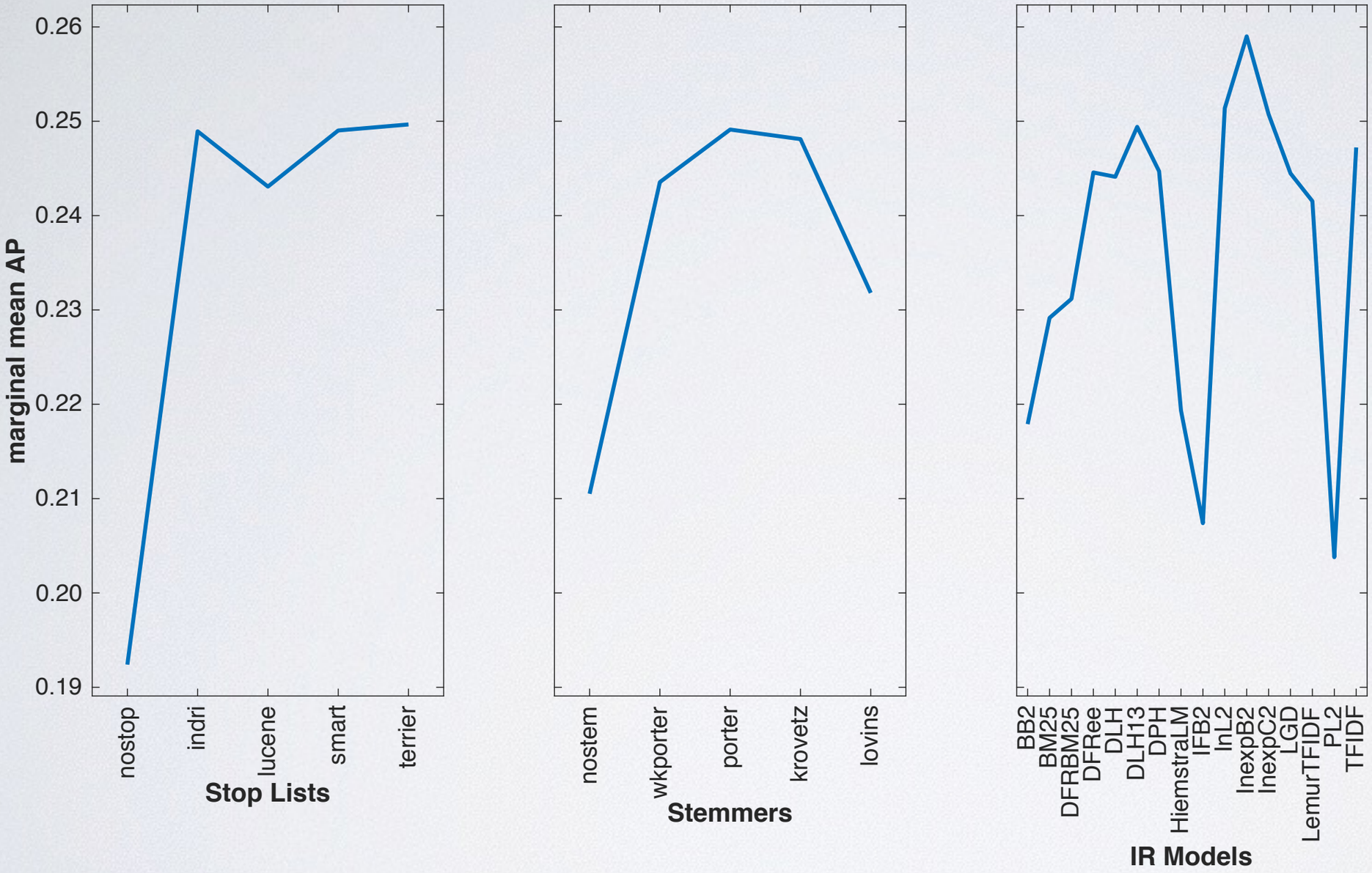
Three Factors: Strength of Association

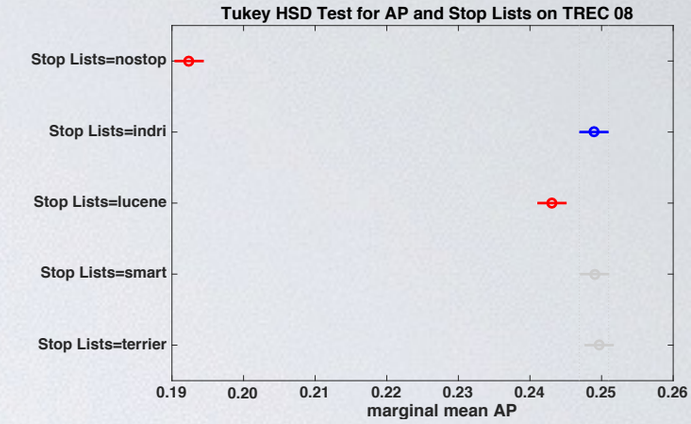
Collection	LUG	Effects	AP	P@10	RBP	nDCG@20	ERR@20
TREC 08	Stemmers	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	**0.0986** (0.00)	**0.0913** (0.00)	**0.1000** (0.00)	**0.1006** (0.00)	**0.0799** (0.00)
		$\hat{\omega}^2_{\langle \text{Stemmers} \rangle}$	0.0439 (0.00)	0.0165 (0.00)	0.0190 (0.00)	0.0268 (0.00)	0.0071 (0.00)
		$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0535 (0.00)	**0.0615** (0.00)	**0.0666** (0.00)	**0.0707** (0.00)	0.0521 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \rangle}$	-0.0003 (0.83)	-0.0005 (0.98)	-0.0005 (0.98)	-0.0006 (0.99)	-0.0004 (0.95)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	0.1565 (0.00)	0.1765 (0.00)	0.1969 (0.00)	**0.2006** (0.00)	**0.1622** (0.00)
		$\hat{\omega}^2_{\langle \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0022 (1.00)	-0.0014 (0.99)	-0.0020 (1.00)	-0.0018 (1.00)	-0.0016 (0.99)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{Stemmers} \times \text{IR Models} \rangle}$	-0.0111 (1.00)	-0.0105 (1.00)	-0.0110 (1.00)	-0.0110 (1.00)	-0.0102 (1.00)
	n-grams	$\hat{\omega}^2_{\langle \text{Stop Lists} \rangle}$	0.0396 (0.00)	0.0423 (0.00)	0.0445 (0.00)	0.0479 (0.00)	0.0304 (0.00)
		$\hat{\omega}^2_{\langle n\text{-grams} \rangle}$	0.0087 (0.00)	0.0031 (0.00)	0.0008 (0.00)	0.0023 (0.00)	0.0093 (0.00)
		$\hat{\omega}^2_{\langle \text{IR Models} \rangle}$	0.0550 (0.00)	0.0545 (0.00)	0.0548 (0.00)	**0.0637** (0.00)	0.0307 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times n\text{-grams} \rangle}$	0.0035 (0.00)	0.0023 (0.00)	0.0024 (0.00)	0.0029 (0.00)	0.0032 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times \text{IR Models} \rangle}$	**0.0928** (0.00)	**0.1129** (0.00)	**0.1231** (0.00)	**0.1277** (0.00)	**0.0940** (0.00)
		$\hat{\omega}^2_{\langle n\text{-grams} \times \text{IR Models} \rangle}$	0.0080 (0.00)	0.0050 (0.00)	0.0059 (0.00)	0.0050 (0.00)	0.0040 (0.00)
		$\hat{\omega}^2_{\langle \text{Stop Lists} \times n\text{-grams} \times \text{IR Models} \rangle}$	-0.0038 (0.99)	-0.0040 (0.99)	-0.0032 (0.99)	-0.0034 (0.99)	-0.0028 (0.99)

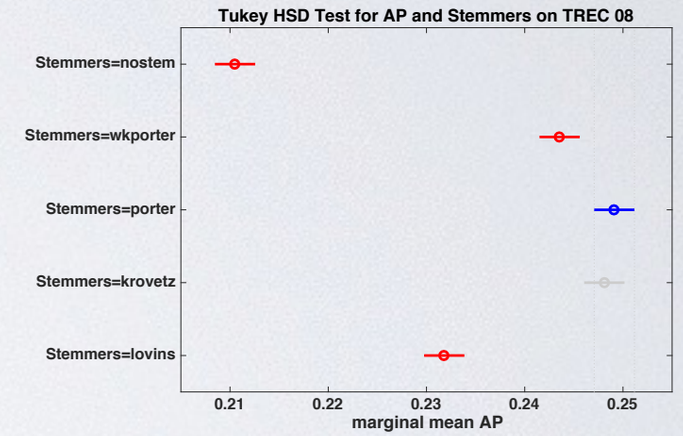
Importance of **linguistic resources**

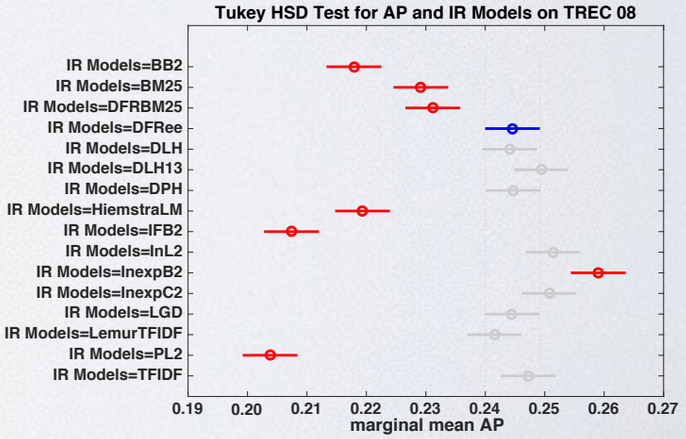
- For the stemmer group, the stop list has always a higher SoA than the IR model and the stemmer and the stop list have a medium effect size
  - The stemmer*model interaction effect which is never significant
- N-grams tend to reduce the stop list effect and to increase the IR model one
  - The n-grams*model interaction effect which is small but statistically significant
- The stop list*model interaction effect is always the biggest of effects (main and interaction ones)
- Not all the measures detect similarly well components effects, e.g. ERR@20 almost ignores the stemmer

Three Factors: Main Effects



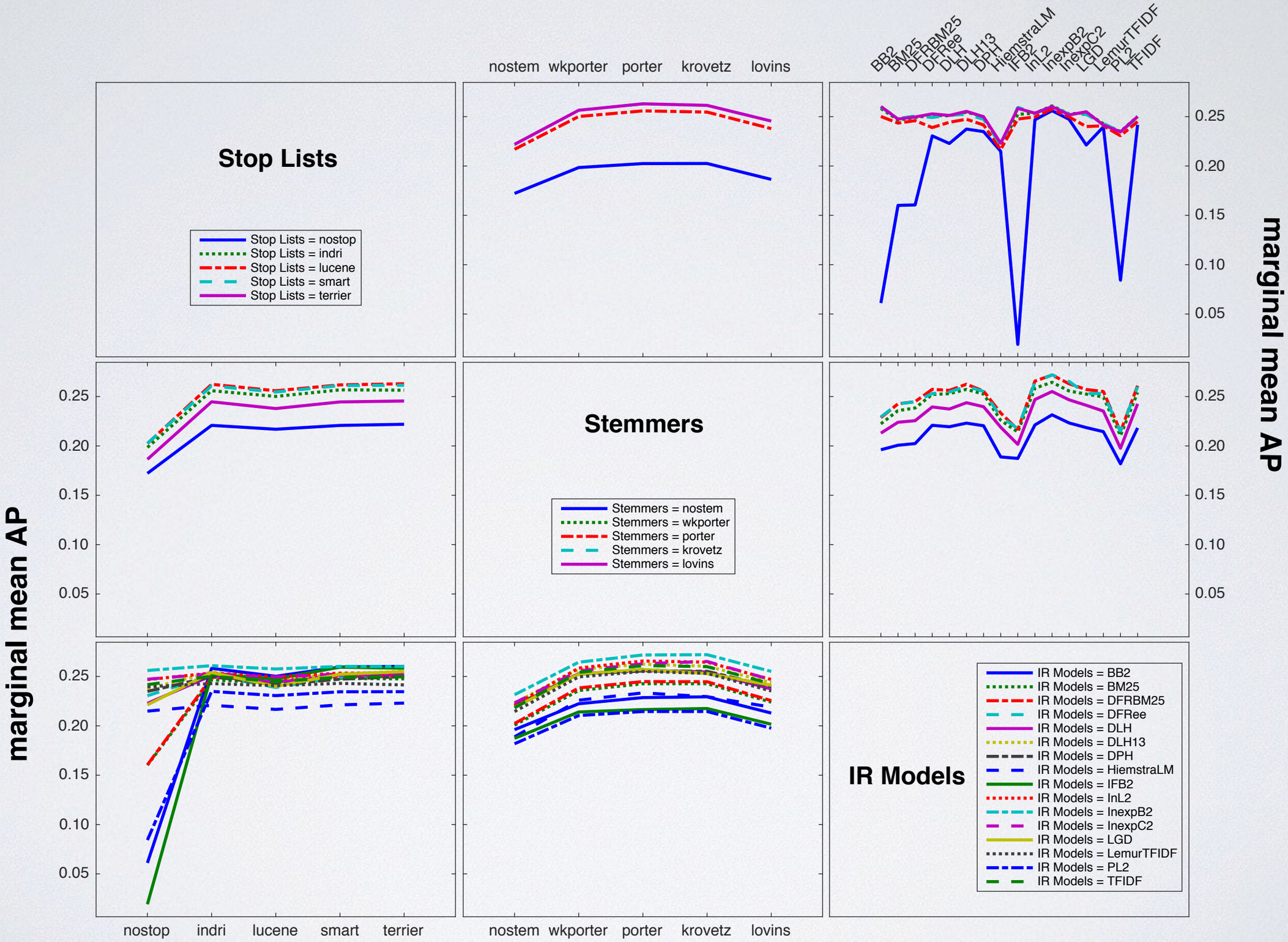




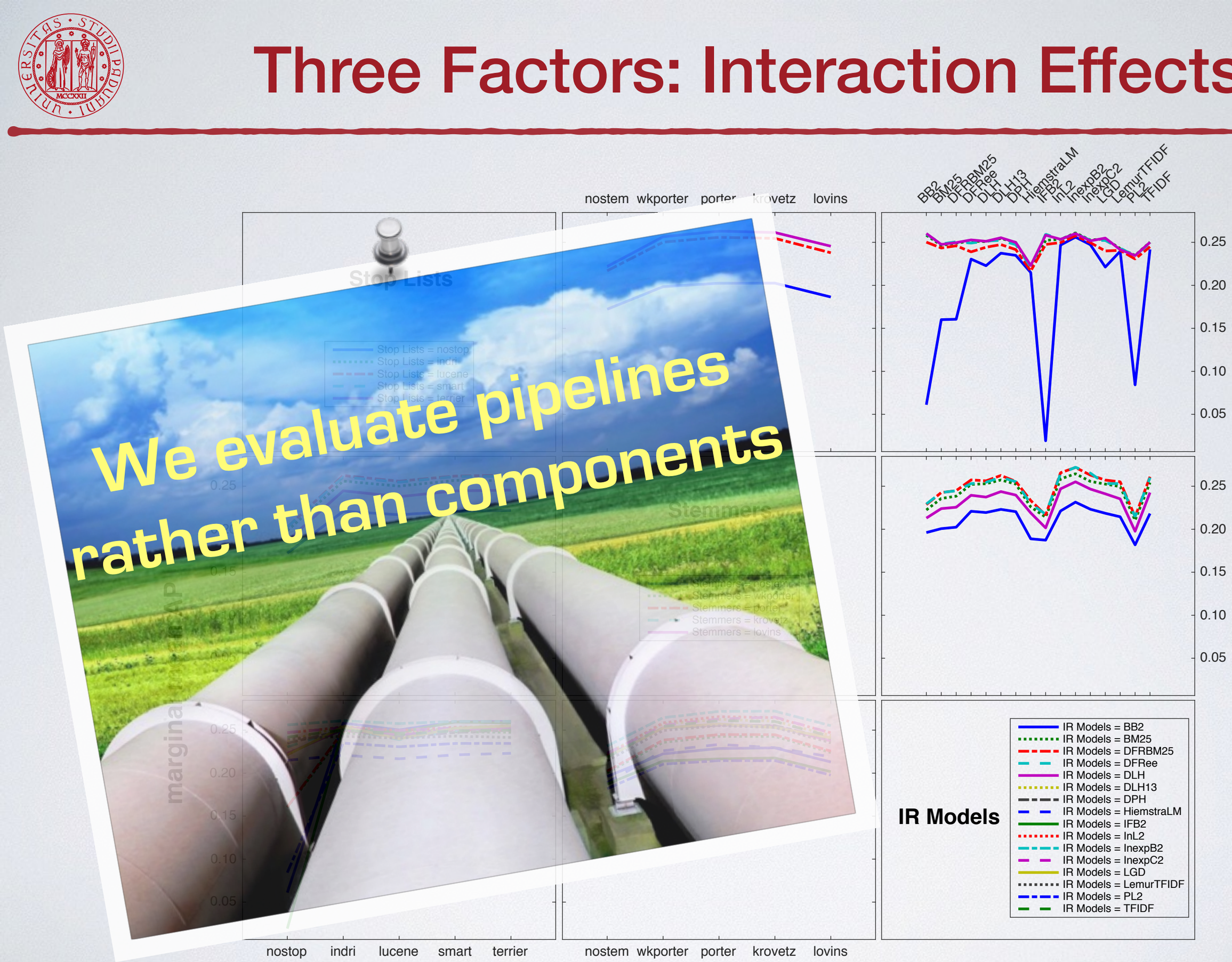


Main Effects of Stop Lists, Stemmers, and IR Models for AP on collection TREC 08

Three Factors: Interaction Effects



Three Factors: Interaction Effects



Wrap up

Summary

- We developed a methodology based on GLMM to break down components effects in a grid of points
  - The most prominent effects are those of stop lists and IR models, as well as their interactions, while stemmers and n-grams play a smaller role
  - Stemmers produce more variation on system performances than n-grams. Overall, this highlights importance of linguistic resources
  - Measures explain system and component effects differently one from the other and not all the measures seem to be suitable for all the cases as it happens for ERR@20 which almost does not detect the stemmer effect
- These insights can be useful to understand where to invest effort and resources for improving components, since they give us an idea of the actual impact of a family of components on the overall performances

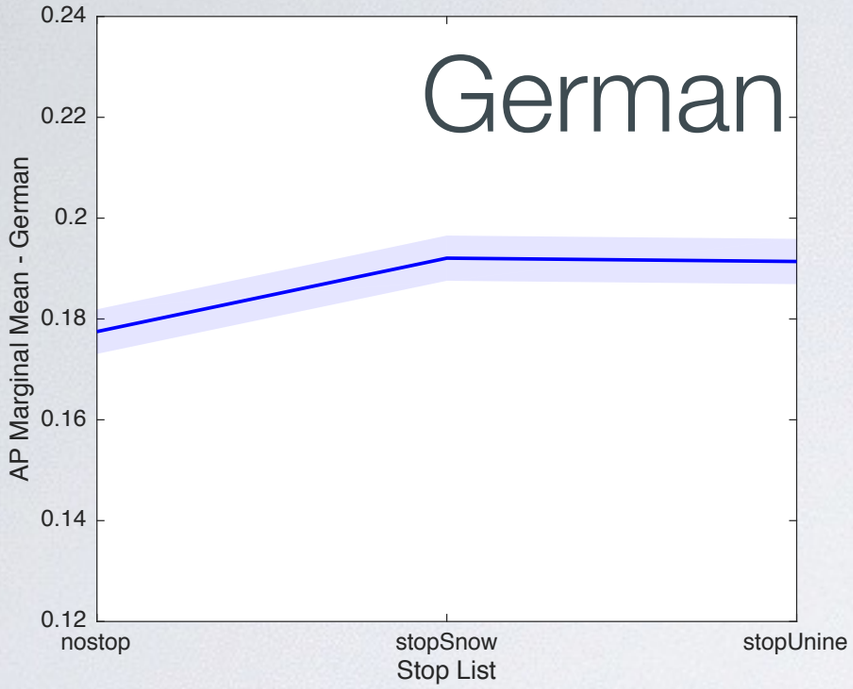
Ahead of Us

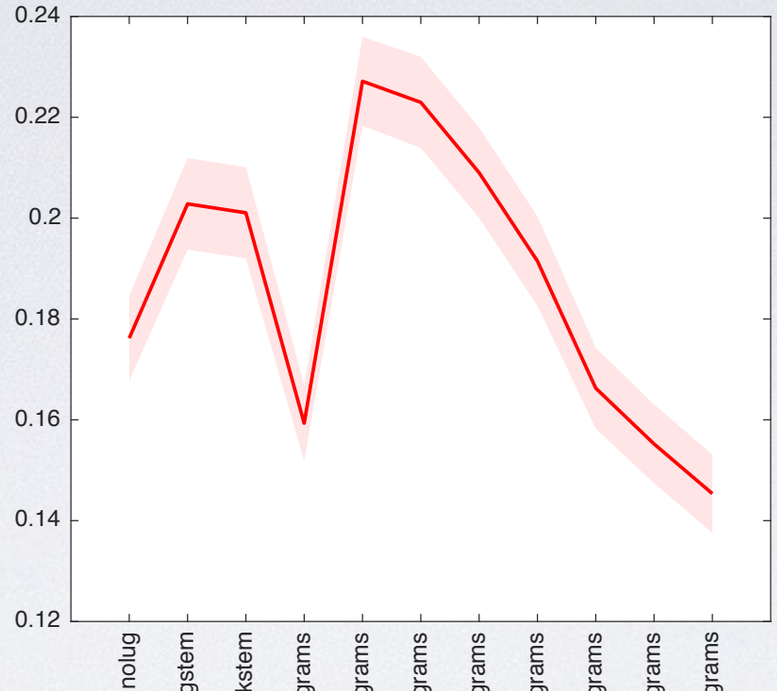
- We plan to further investigate the impact of the measures on the determination of effect sizes

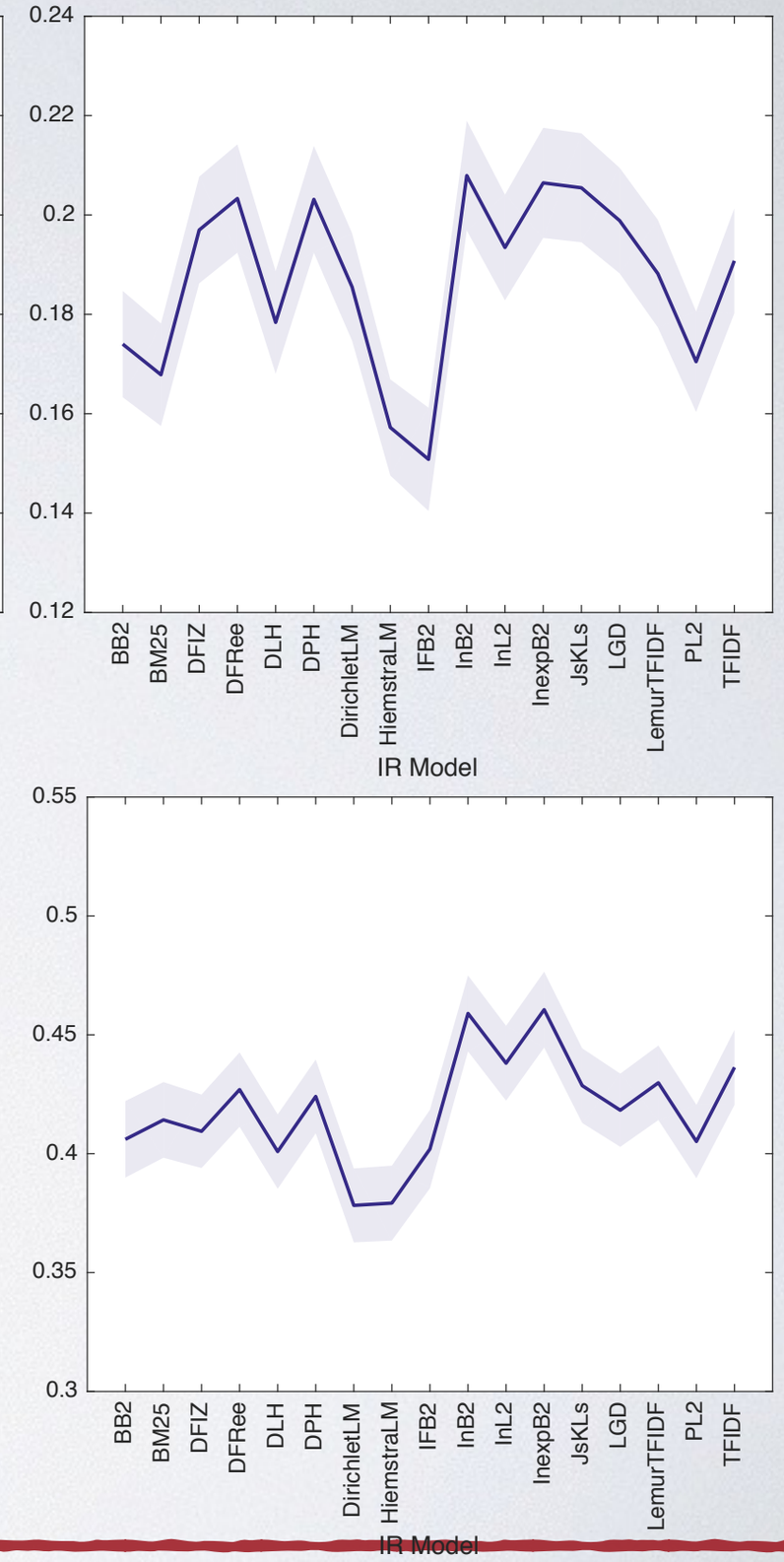
- We intend to apply this kind of analyses in the case of multiple languages, e.g. on CLEF data, in order to study the language effect

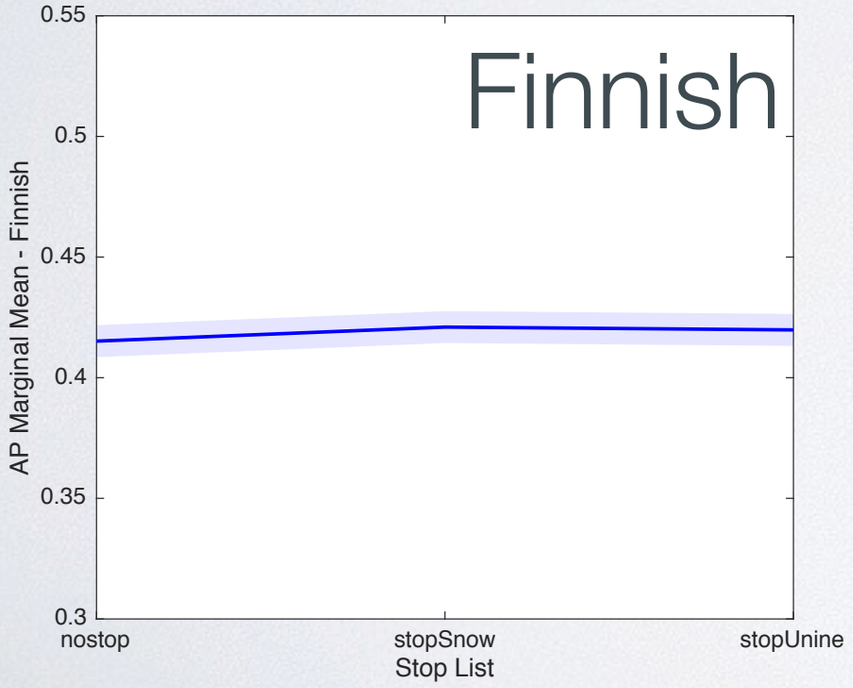
- Open challenges concern how to assess the topic/system interaction effects and how to extend this methodology to data from system participating in evaluation campaigns

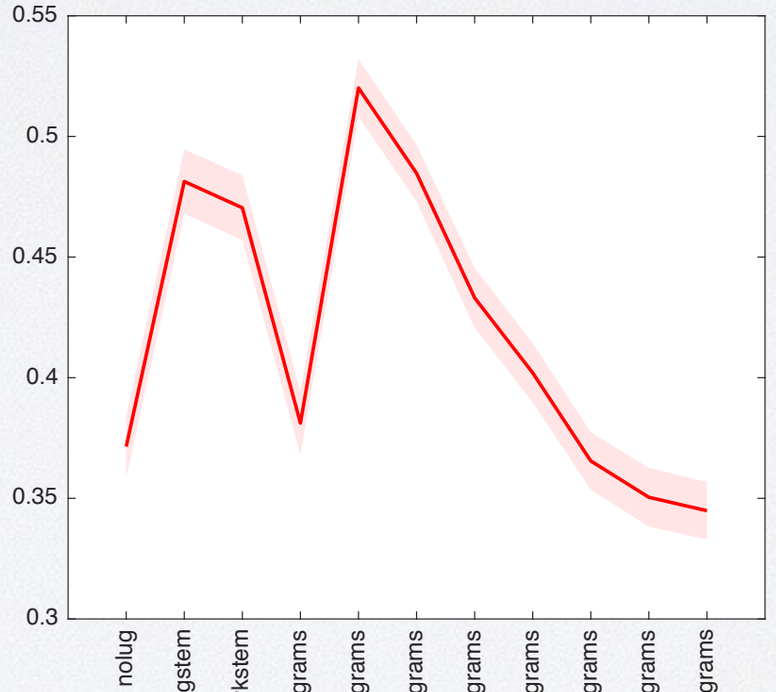
CLEF 2003: Main Effects for Some Languages



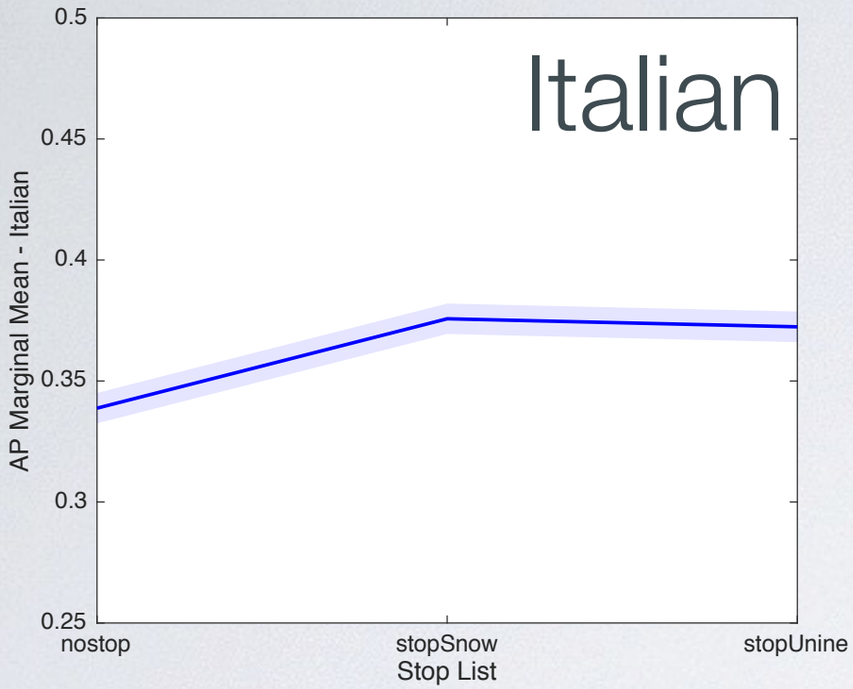


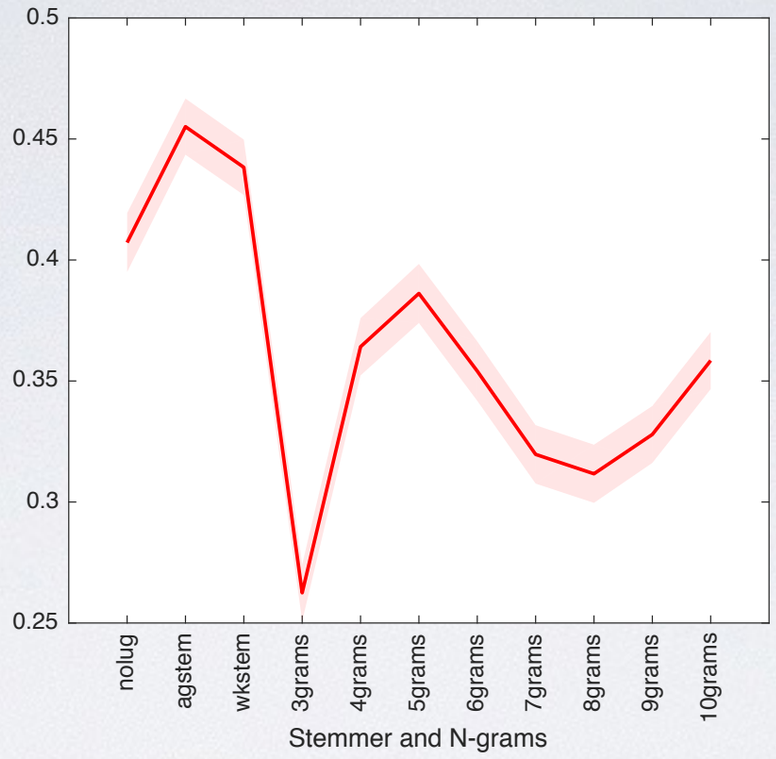


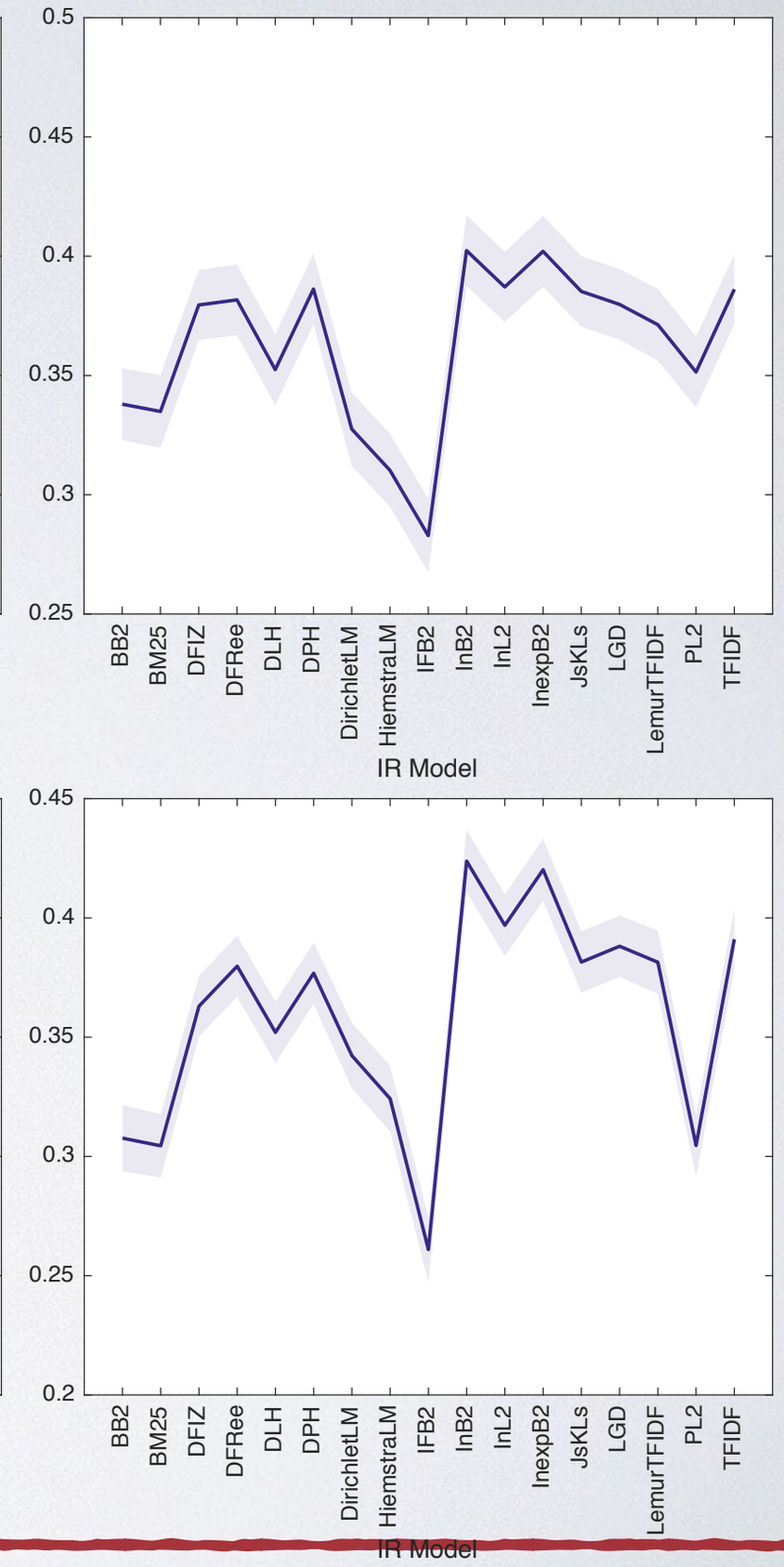


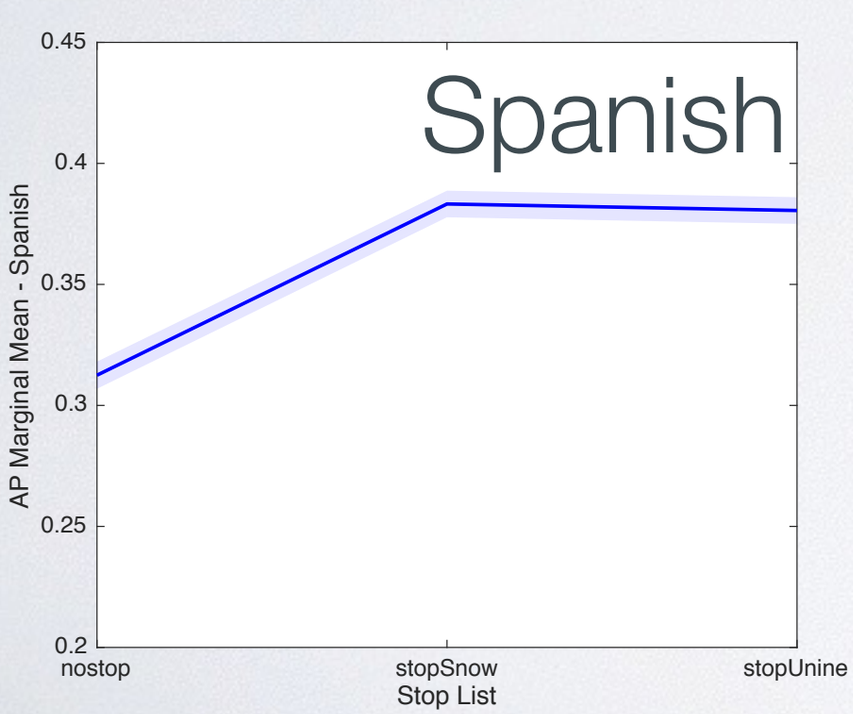


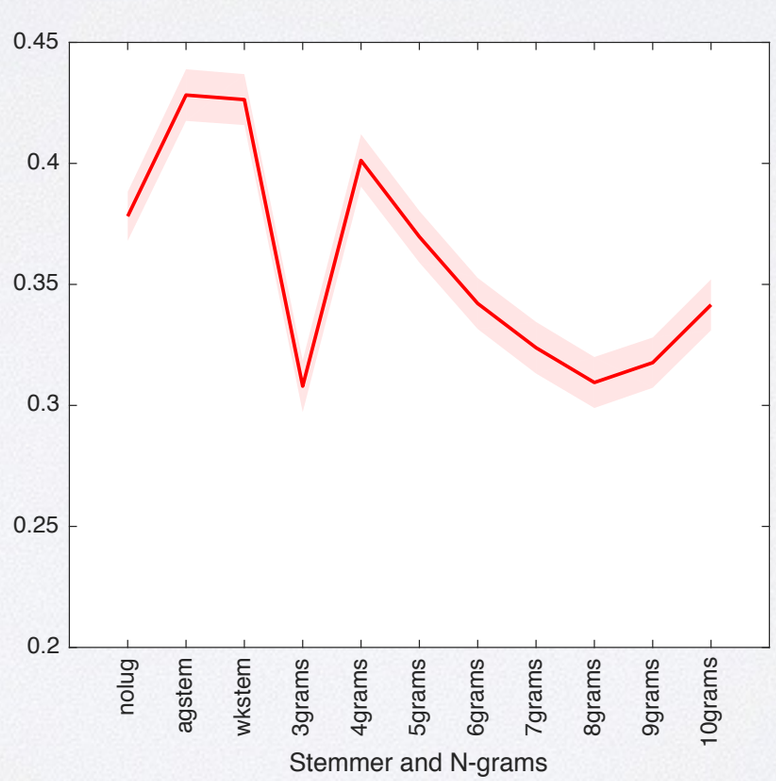
CLEF 2003: Main Effects for Some Languages



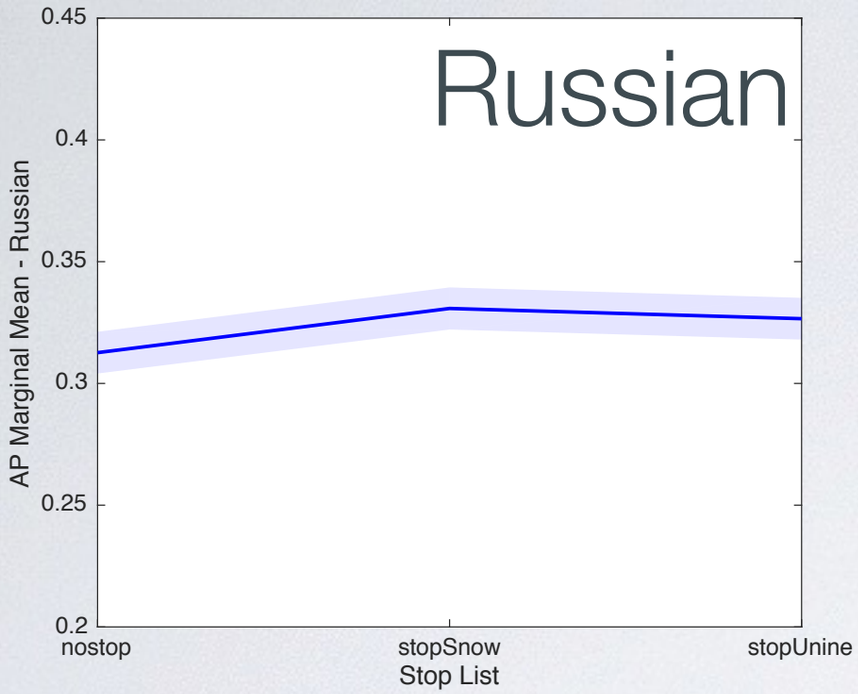


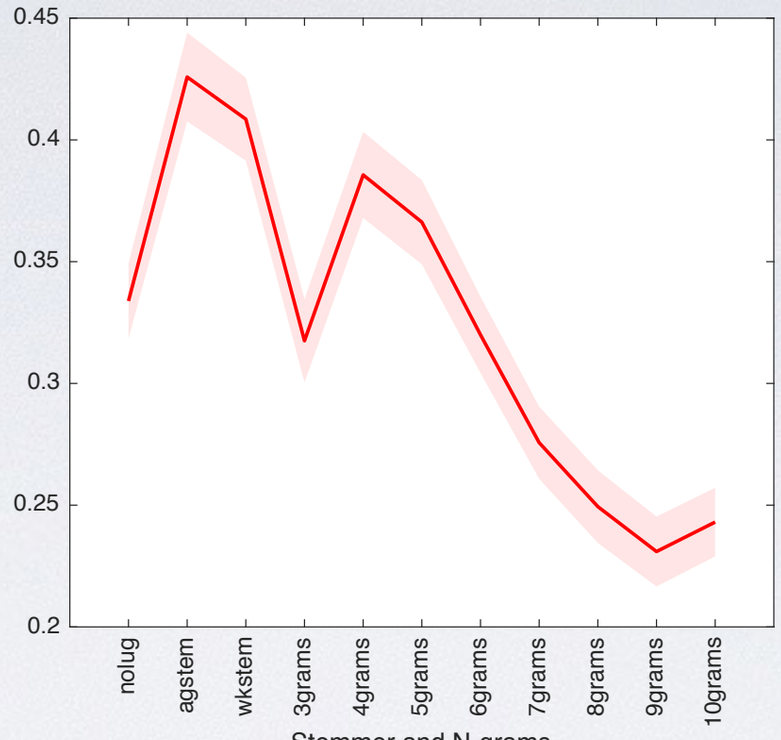


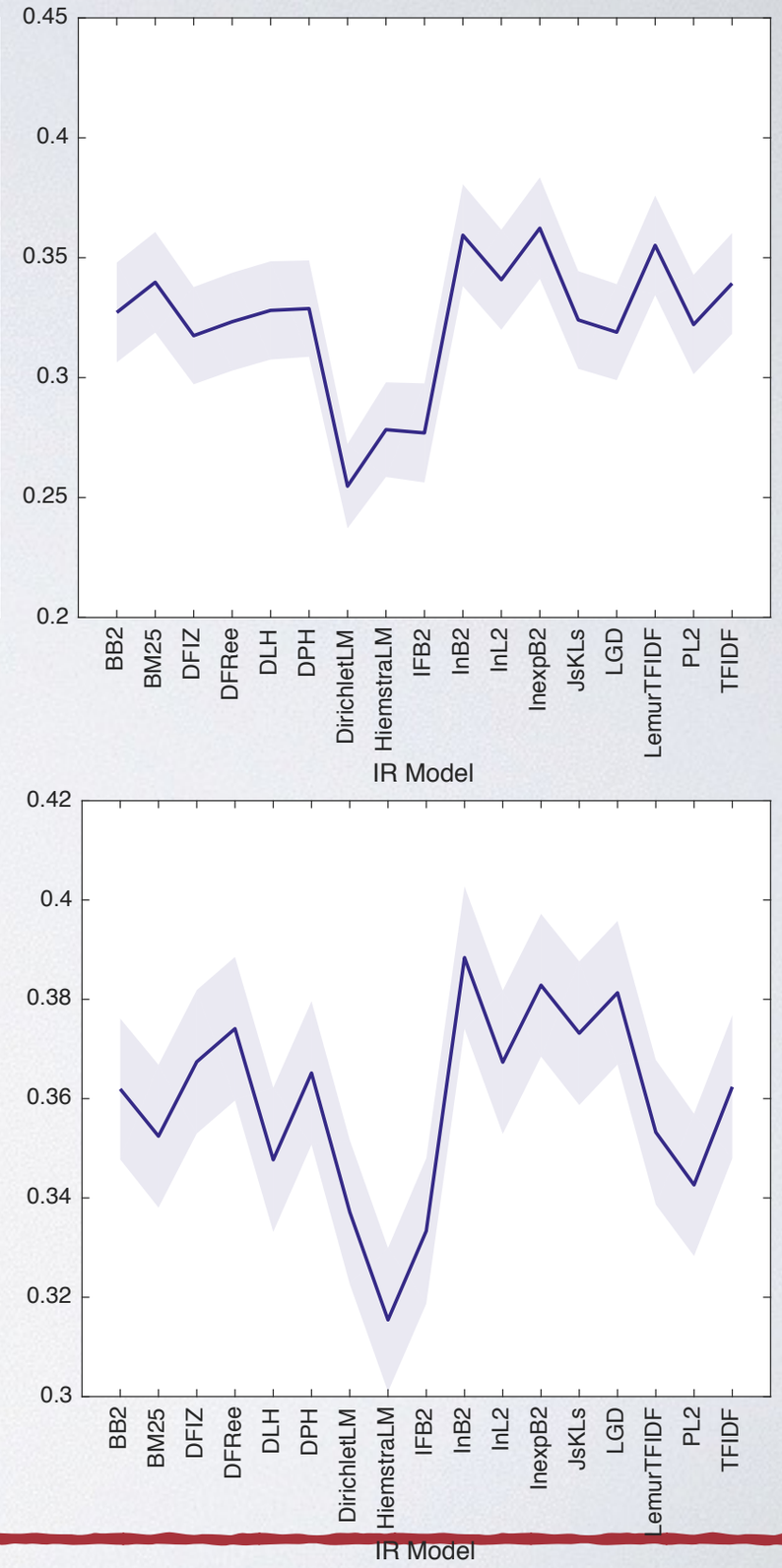


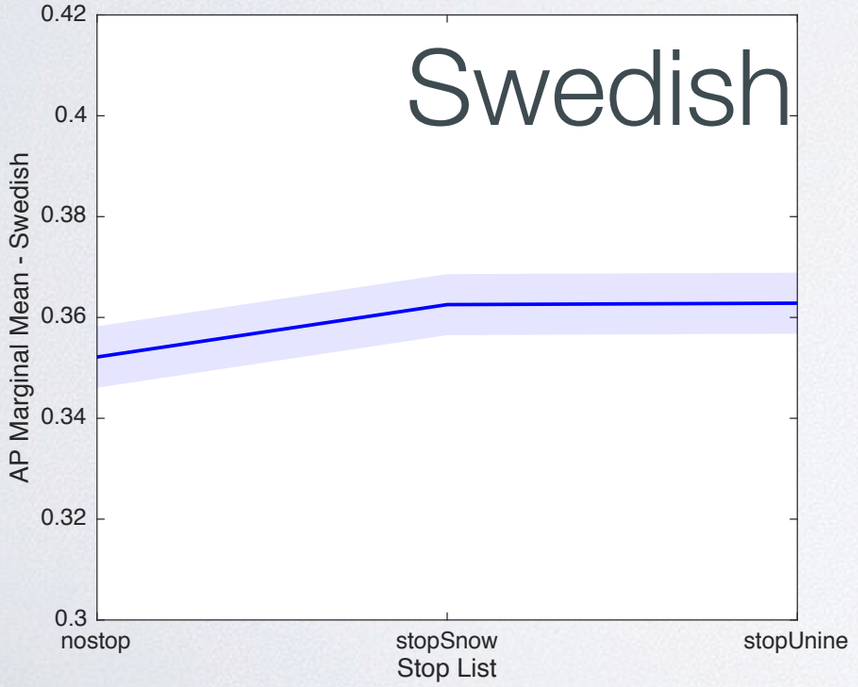


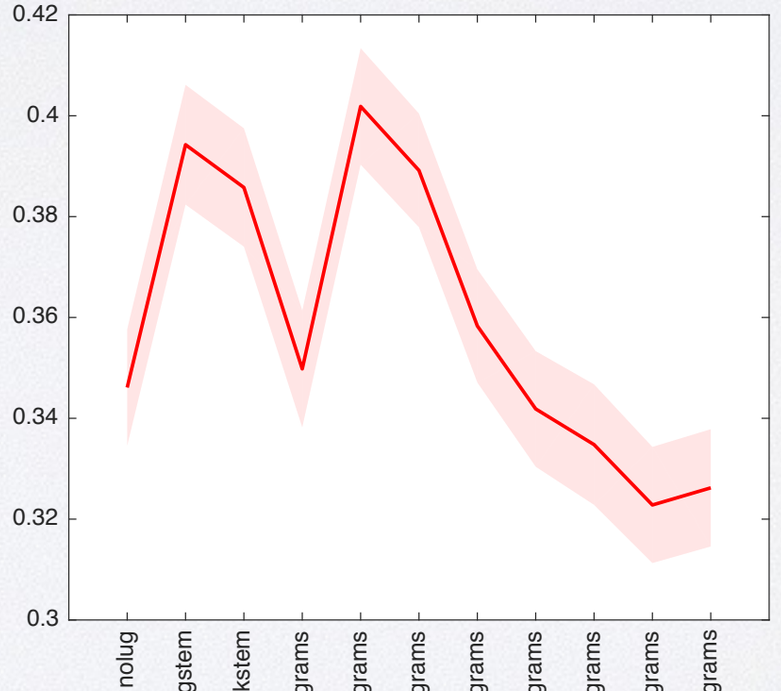
CLEF 2003: Main Effects for Some Languages









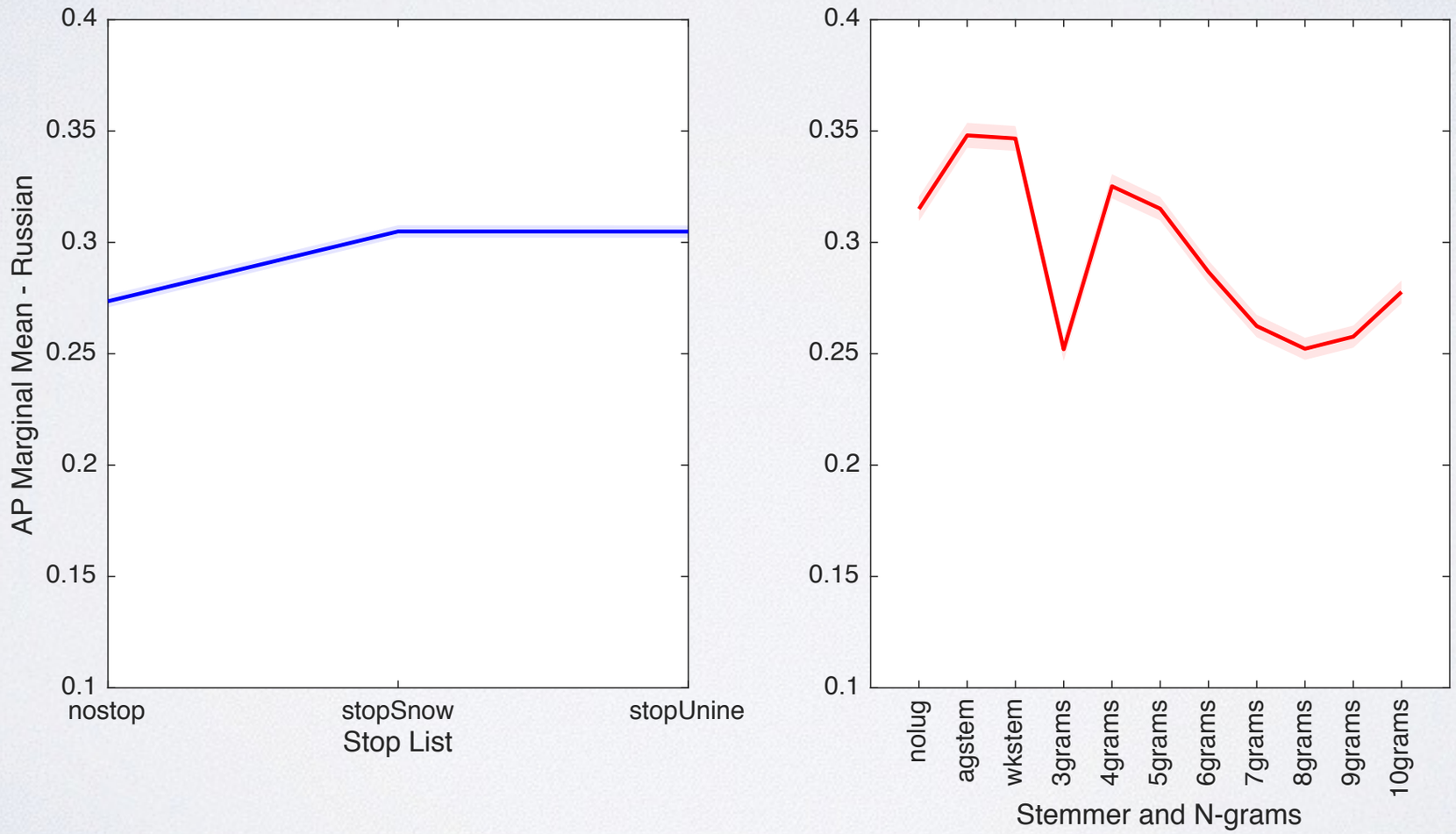


Going Multilingual… Not so easy

- Components are much sparser
  - less fine-grained GoPs
- Linguistic processing may differ a lot
  - tokenization
- Not all the components make sense in all the languages
  - decompounding
- What does make components “equivalent” across languages?

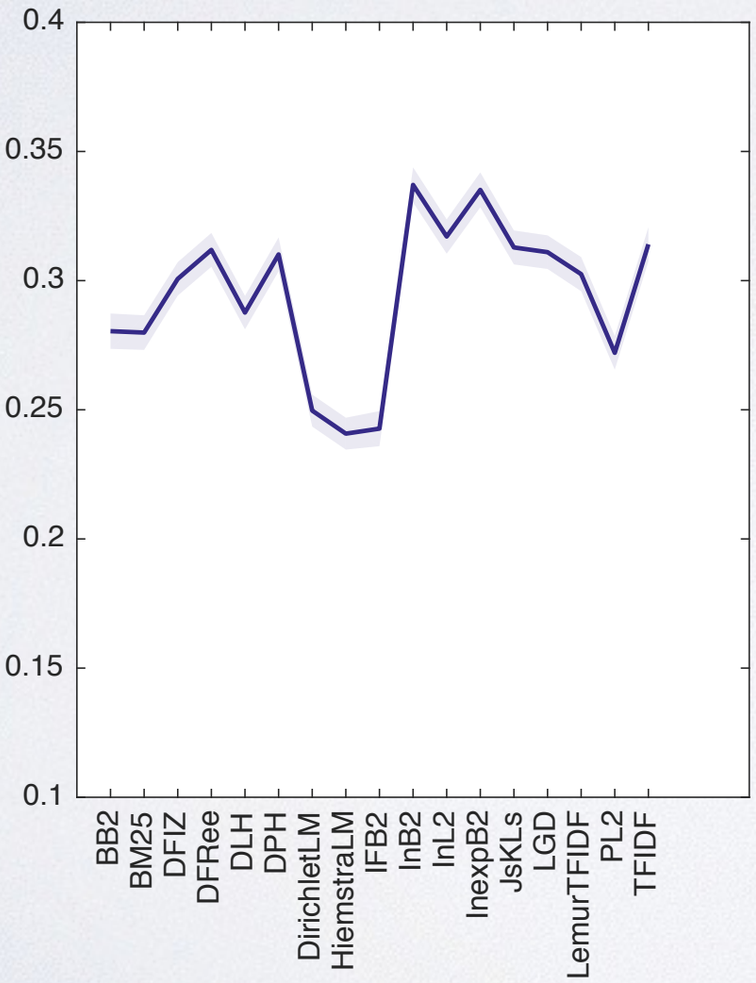
Four Factors Model: CLEF 2003 Main Effects

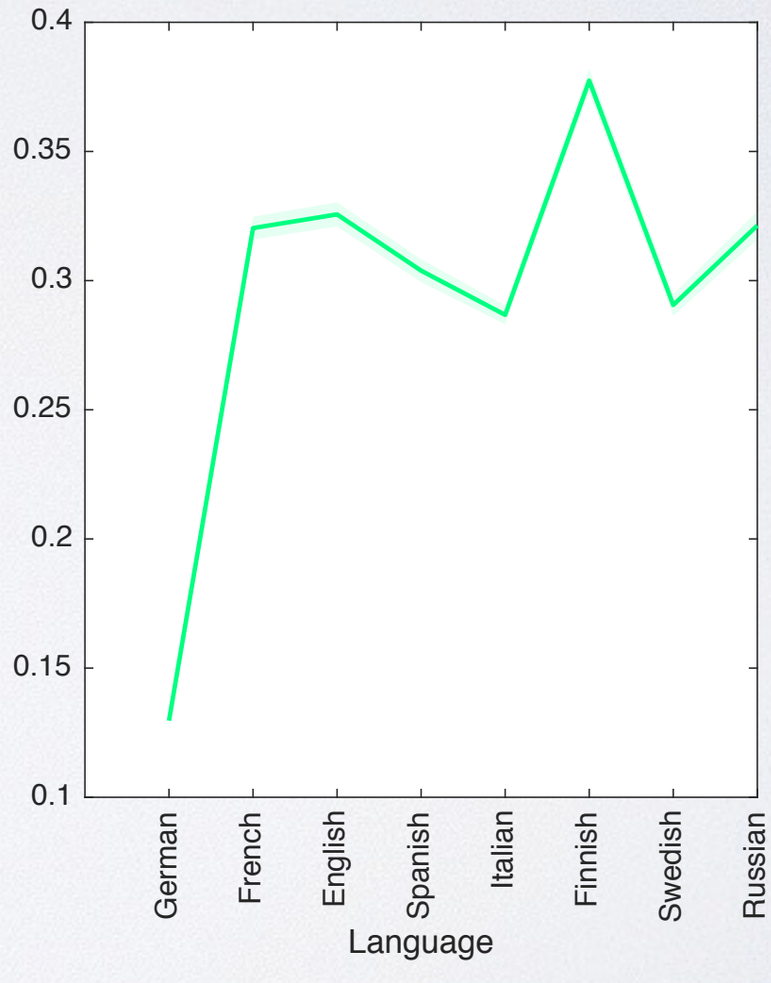
$$Y_{ijklm} = \underbrace{\mu..... + \tau_i + \alpha_j + \beta_k + \gamma_l + \delta_m}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \alpha\delta_{jm} + \beta\gamma_{kl} + \beta\delta_{km} + \gamma\delta_{lm}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijklm}}_{\text{Error}}$$



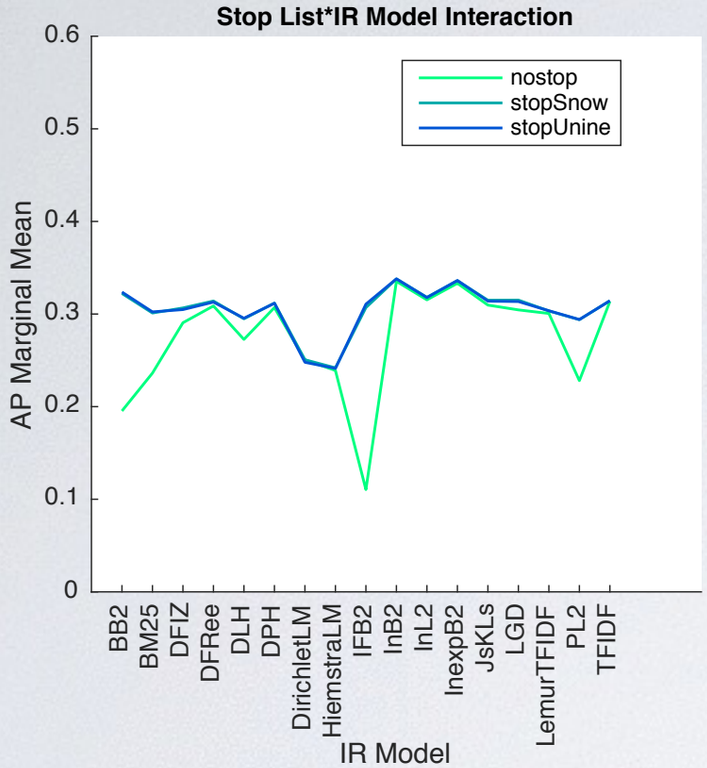
Four Factors Model: CLEF 2003 Main Effects

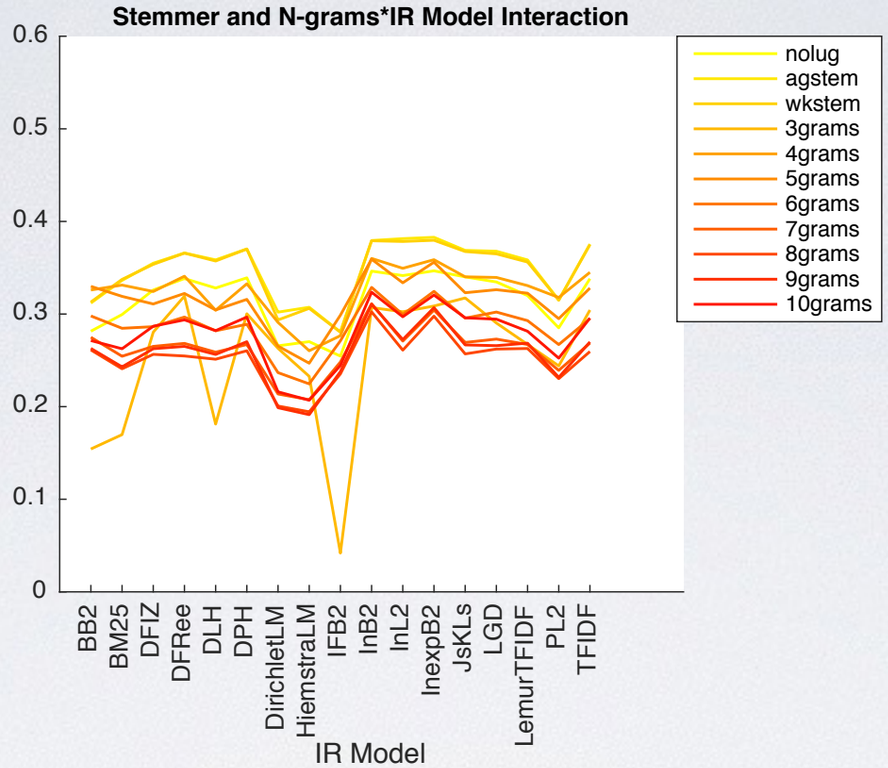
$$Y_{ijklm} = \underbrace{\mu..... + \tau_i + \alpha_j + \beta_k + \gamma_l + \delta_m}_{\text{Main Effects}} + \underbrace{\alpha\beta_{jk} + \alpha\gamma_{jl} + \alpha\delta_{jm} + \beta\gamma_{kl} + \beta\delta_{km} + \gamma\delta_{lm}}_{\text{Interaction Effects}} + \underbrace{\varepsilon_{ijklm}}_{\text{Error}}$$

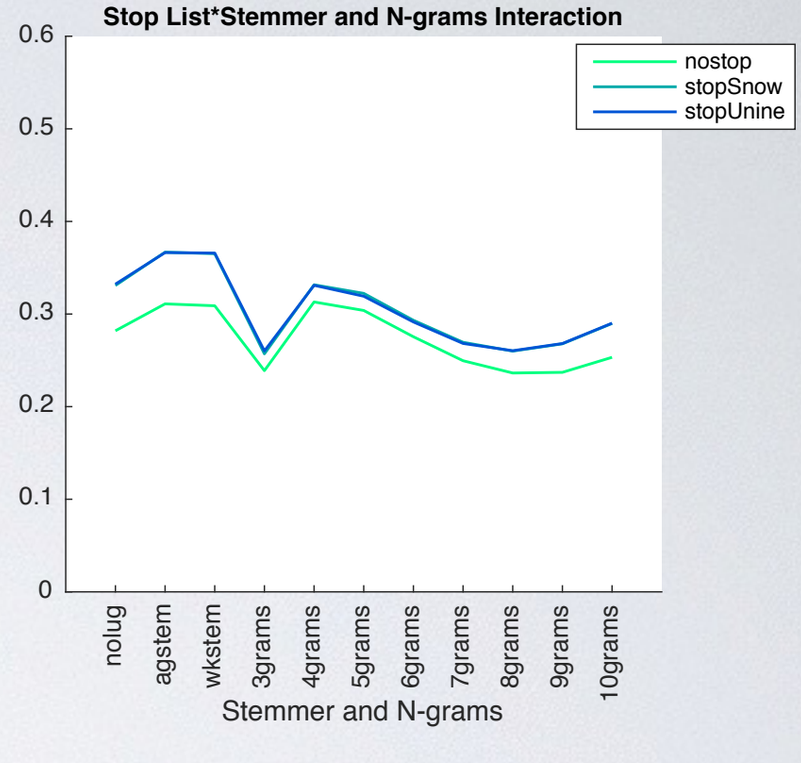


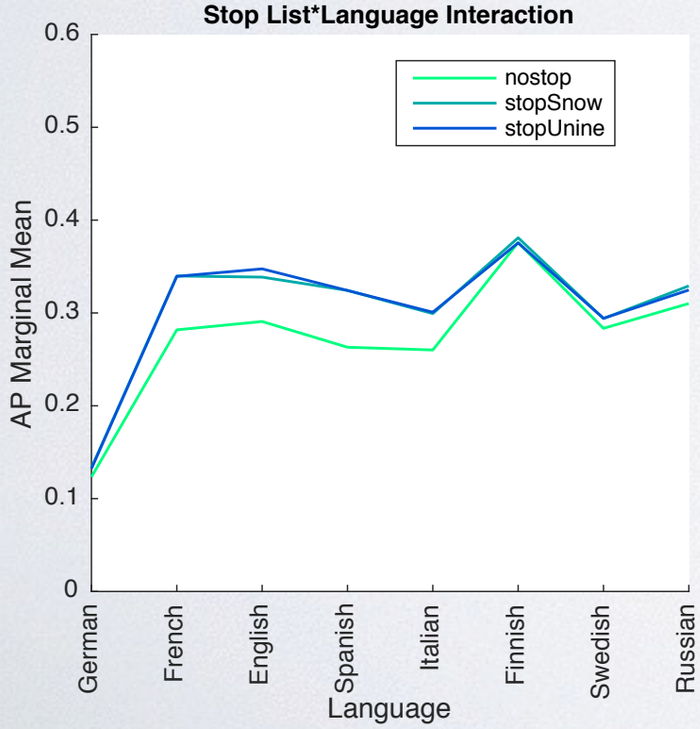


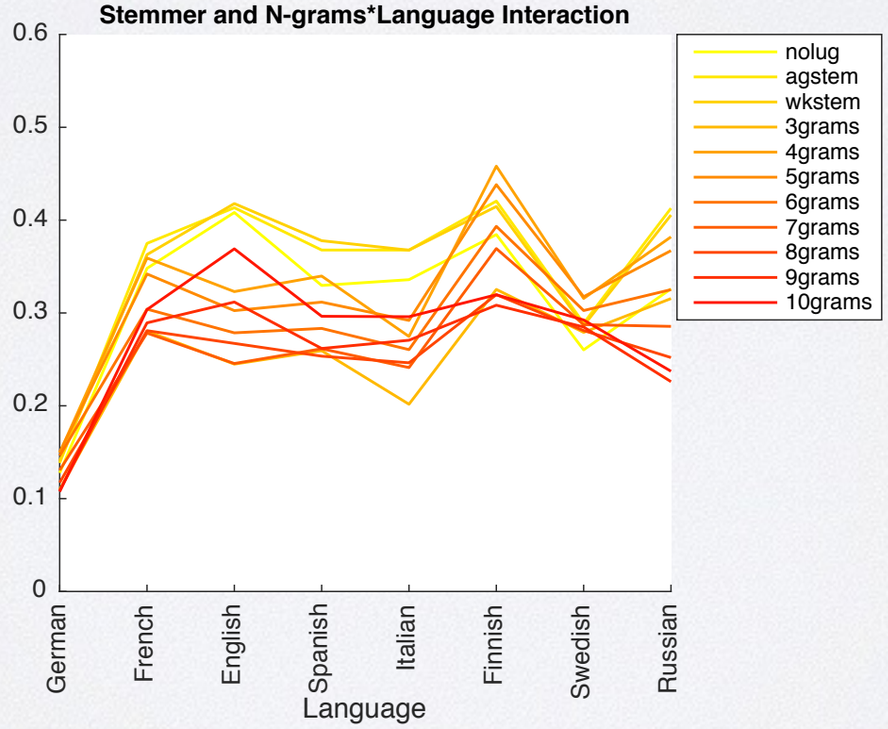
Four Factors Model: CLEF 2003 Interaction Effects

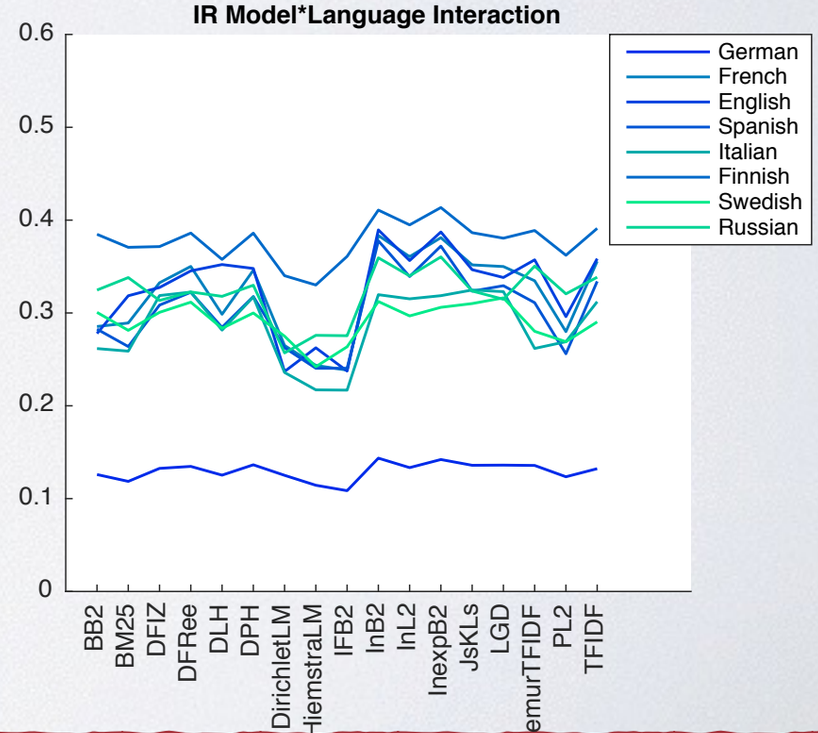












Take Home Message

References (1/2)

- Arguello, J., Crane, M., Diaz, F., Lin, J., and Trotman, A. (2015). Report on the SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR). *SIGIR Forum*, 49(2):107-116.
- Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, Cambridge (MA), USA.
- Ferro, N. and Harman, D. (2010). CLEF 2009: Grid@CLEF Pilot Track Overview. In Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peas, A., and Roda, G., editors, *Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments - Tenth Workshop of the Cross-Language Evaluation Forum (CLEF 2009). Revised Selected Papers*, pages 552-565. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany.
- Ferro, N. and Silvello, G. (2016). A General Linear Mixed Models Approach to Study System Component Effects. In Perego, R., Sebastiani, F., Aslam, J., Ruthven, I., and Zobel, J., editors, *Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016)*, pages 25–34. ACM Press, New York, USA.
- Ferro, N. and Silvello, G. (2016b). The CLEF Monolingual Grid of Points. In Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Inter- action. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016)*, pages 16–27. Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany.

References (2/2)

- Fuhr, N. (2010). IR between Science and Engineering, and the Role of Experimentation. In Agosti, M., Ferro, N., Peters, C., de Rijke, M., and Smeaton, A., editors, *Multilingual and Multimodal Information Access Evaluation. Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010)*, page 1. Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany.
- Fuhr, N. (2012). Salton Award Lecture: Information Retrieval As Engineering Science. SIGIR Forum, 46(2):19–28.
- Robertson, S. E. (1981). The methodology of information retrieval experiment. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 9-31. Butterworths, London, United Kingdom.
- Tague-Sutcliffe, J. M. and Blustein, J. (1994). A Statistical Analysis of the TREC-3 Data. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385-398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.


ANY
QUESTIONS
?