Comparing weighting models for monolingual information retrieval

Gianni Amati, Claudio Carpineto, and Gianni Romano

Fondazione Ugo Bordoni

Roma

romano@fub.it

Overview

- Three weighting models
- Retrieval feedback
- Experimental settings
- Results
- Conclusions

Document ranking

$$Sim(q,d) = \sum_{t \in q \cap d} w_{t,q} \cdot w_{t,d}$$

q query
d document
t term

$w_{t,q}$ *query term weight*
$w_{t,d}$ *doc term weight*

Okapi

$$w_{t,q} = \frac{(k_3 + 1) \cdot f_{t,q}}{k_3 + f_{t,q}} \cdot \log_2 \frac{D - n_t + 0.5}{n_t + 0.5}$$

$$w_{t,d} = \frac{(k_1 + 1) \cdot f_{t,d}}{k_1 \cdot \left\{ (1 - b) + b \cdot \frac{W_d}{avr\_W_d} \right\} + f_t}$$

Statistical Language Modeling (SLR)

$$w_{t,q} = f_{t,q}$$

$$w_{t,d} =$$

$$\log_2 \frac{f_{t,d} + \square\square_t}{W_d + \square} + \log_2 \frac{\square}{W_d + \square} + \log_2 \square_t$$

$$+ W_q \square \log_2 \frac{\square}{W_d + \square}$$

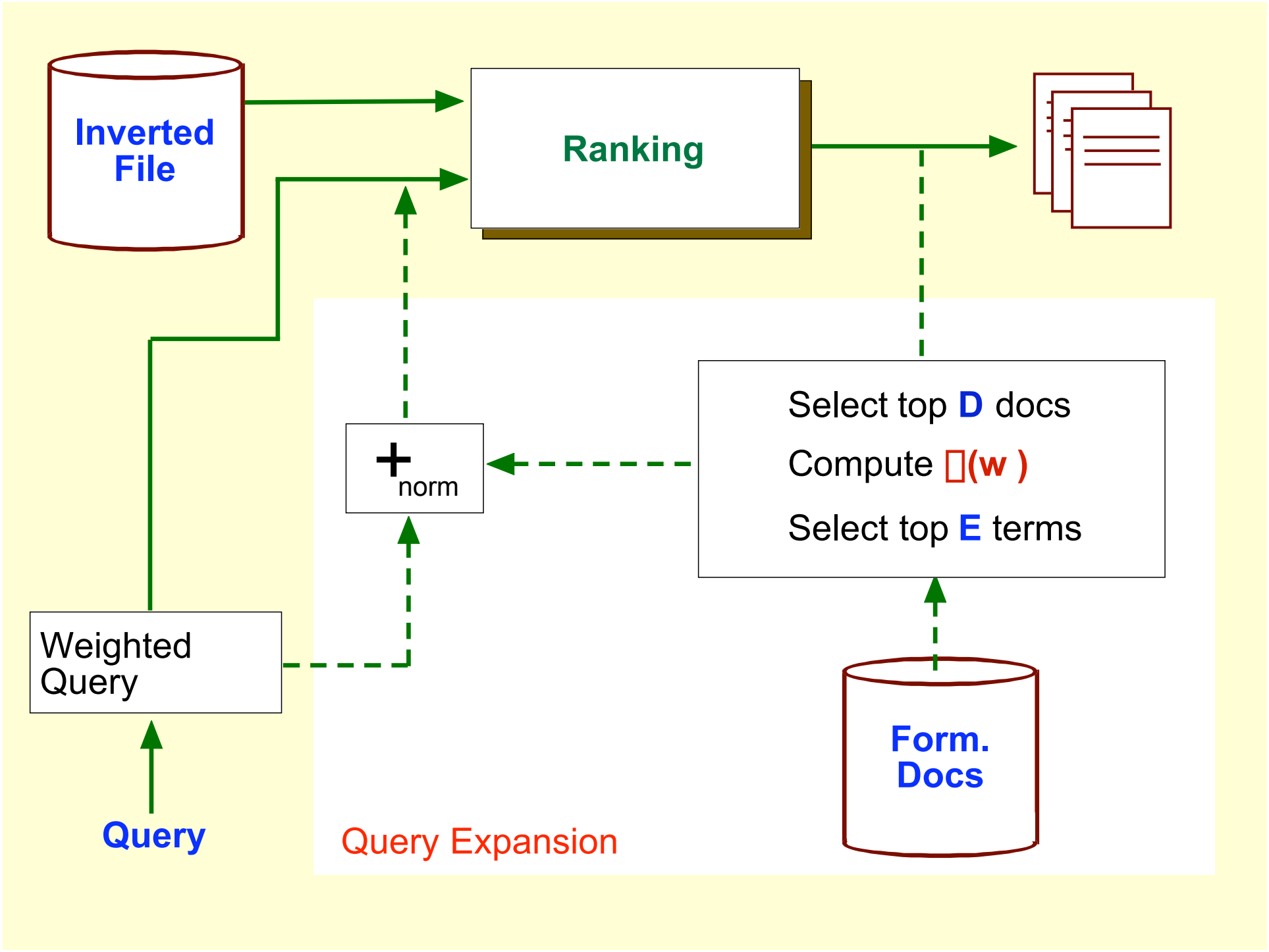
Deviation from randomness (DFR)

$$w_{t,q} = f_{t,q}$$

$$w_{t,d} = \left\{ \log_2 (1 + \lambda_t) + f^*_{t,d} \cdot \log_2 \frac{1 + \lambda_t}{\lambda_t} \right\}$$

$$\lambda_t = \frac{f_t + 1}{n_t \cdot (f^*_{t,d} + 1)}$$

$$f^*_{t,d} = f_{t,d} \cdot \log_2 \left(1 + \frac{c \cdot avr\_W_d}{W_d}\right)$$


Inverted File
Ranking
Select top D docs
Compute □(w)
Select top E terms
+ norm
Weighted Query
Query
Form. Docs
Query Expansion

Document ranking

$$Sim(q,d) = \sum_{t \in q \cap d} w_{t,q} \cdot w_{t,d}$$

q query
d document
t term

$w_{t,q}$ *query term weight*
$w_{t,d}$ *doc term weight*

Retrieval feedback

$$Sim(q_{exp}, d) = \sum_{t \in q\_exp \cap d} w_{t,q\_exp} \cdot w_{t,d}$$

$$w_{t,q\_exp} = \alpha \cdot \frac{w_{t,q}}{max_q\, w_{t,q}} + \beta \cdot \frac{KLD_{t,d}}{max_d\, KLD_{t,d}}$$

$$KLD_{t,d} = f_{t,d} \cdot \log_2 \frac{f_{t,d}}{f_t}$$

Test Collections

French, Italian, Spanish monolingual

Query ***title + description***

Stemming

Porter algorithms ***(snowball.tartarus.org)***

Stop list ***Savoy***

French

	AvPrec	Prec-at-5	Prec-at-10
SLM	0.4753	0.4538	0.3635
SLM+RF	0.4372	0.4192	0.3462
Okapi	0.5030	0.4385	0.3654
Okapi+RF	0.5054	0.4769	0.3942
DFR	0.5116	0.4577	0.3654
DFR+RF	0.5238	0.4885	0.3981

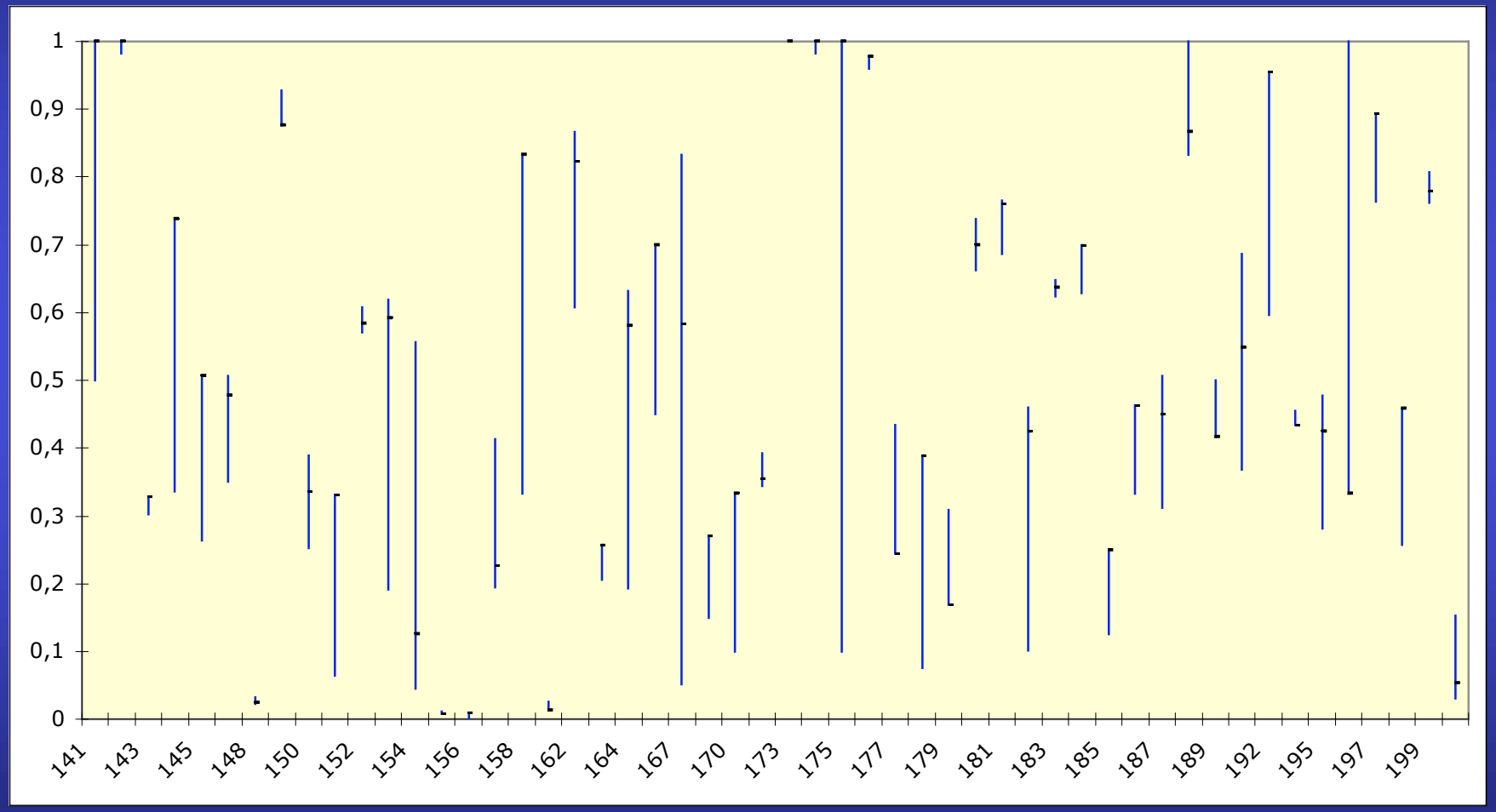
Italian

	AvPrec	Prec-at-5	Prec-at-10
SLM	0.5027	0.4941	0.3824
SLM+RF	0.5095	0.4824	0.3863
Okapi	0.4762	0.4588	0.3510
Okapi+RF	0.5238	0.4824	0.3902
DFR	0.5046	0.4824	0.3725
DFR+RF	0.5364	0.5255	0.4137

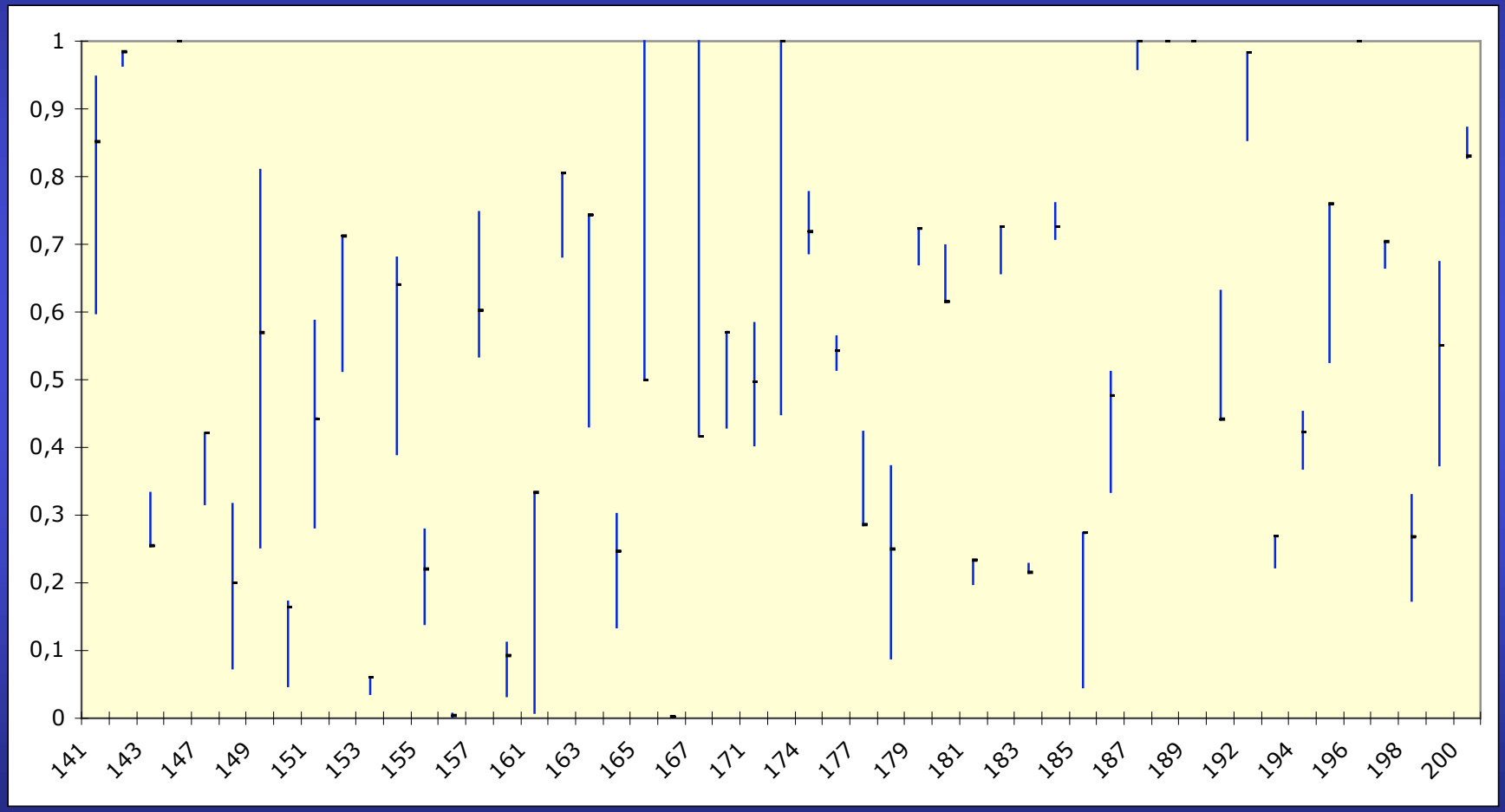
Spanish

	AvPrec	Prec-at-5	Prec-at-10
SLM	0.4720	0.6140	0.5175
SLM+RF	0.5112	0.5825	0.5316
Okapi	0.4606	0.5684	0.5175
Okapi+RF	0.5093	0.6105	0.5491
DFR	0.4907	0.6035	0.5386
DFR+RF	0.5510	0.6140	0.5825

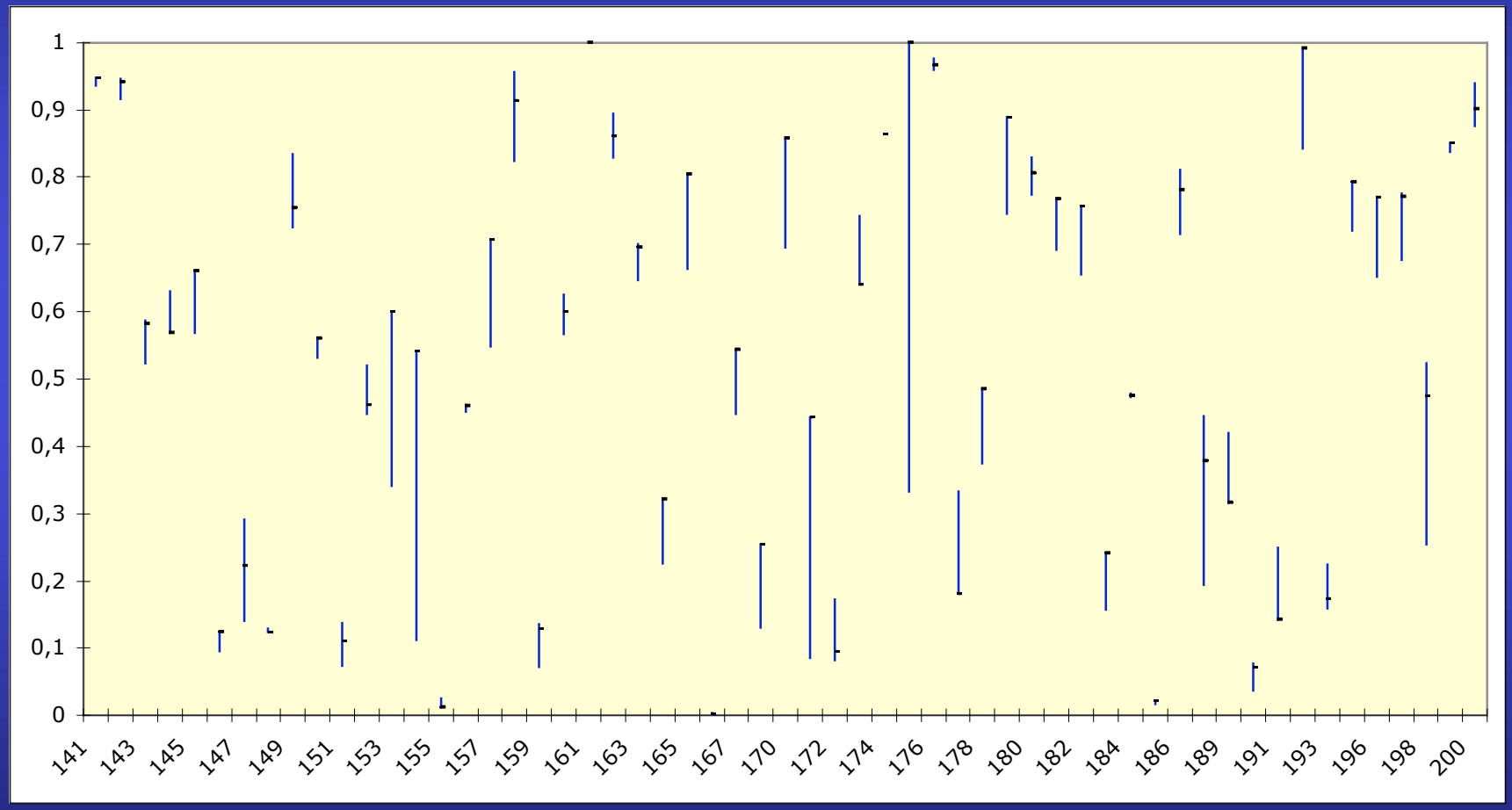
French AvPrec variation



Italian AvPrec variation



Spanish AvPrec variation



Average delta AvPrec

	delta	max	best
French	0.2047	0.5796	0.5238
Italian	0.1596	0.5978	0.5364
Spanish	0.1050	0.5732	0.5510

Ranked performance

	French			Italian			Spanish		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
SLM	11	11	30	10	9	32	16	10	31
Okapi	20	17	15	21	16	14	16	22	19
DFR	21	24	7	20	26	5	25	25	7

Conclusions

- **DFR > Okapi, SLM**
- **Retrieval feedback mostly effective**
- **Performance mostly language independent**

Future experiments with a wide range of factors: query length, model parameters, expansion parameters