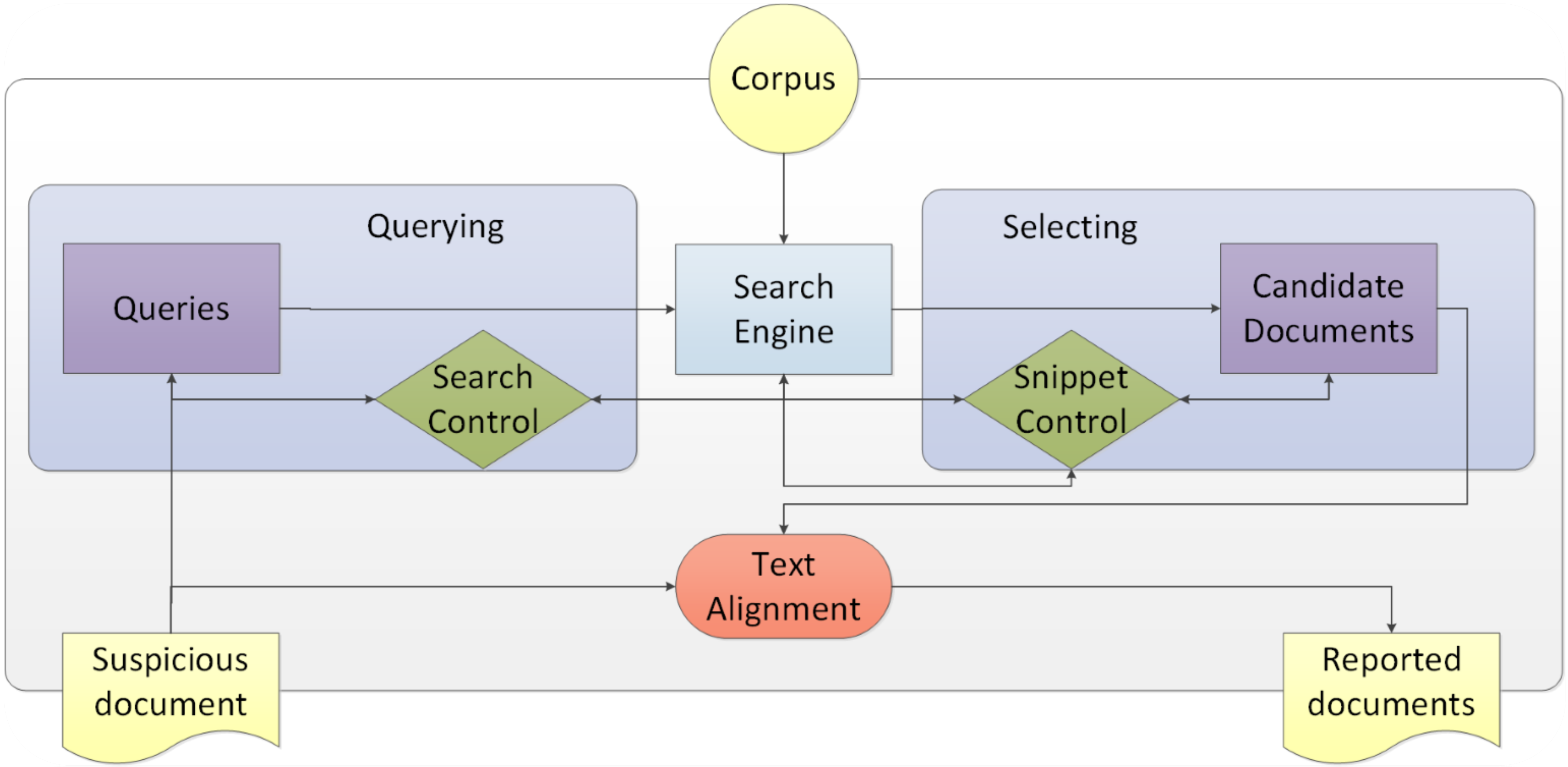
Improving Synoptic Querying for Source Retrieval

Šimon Suchomel

Process Overview



Building of Queries

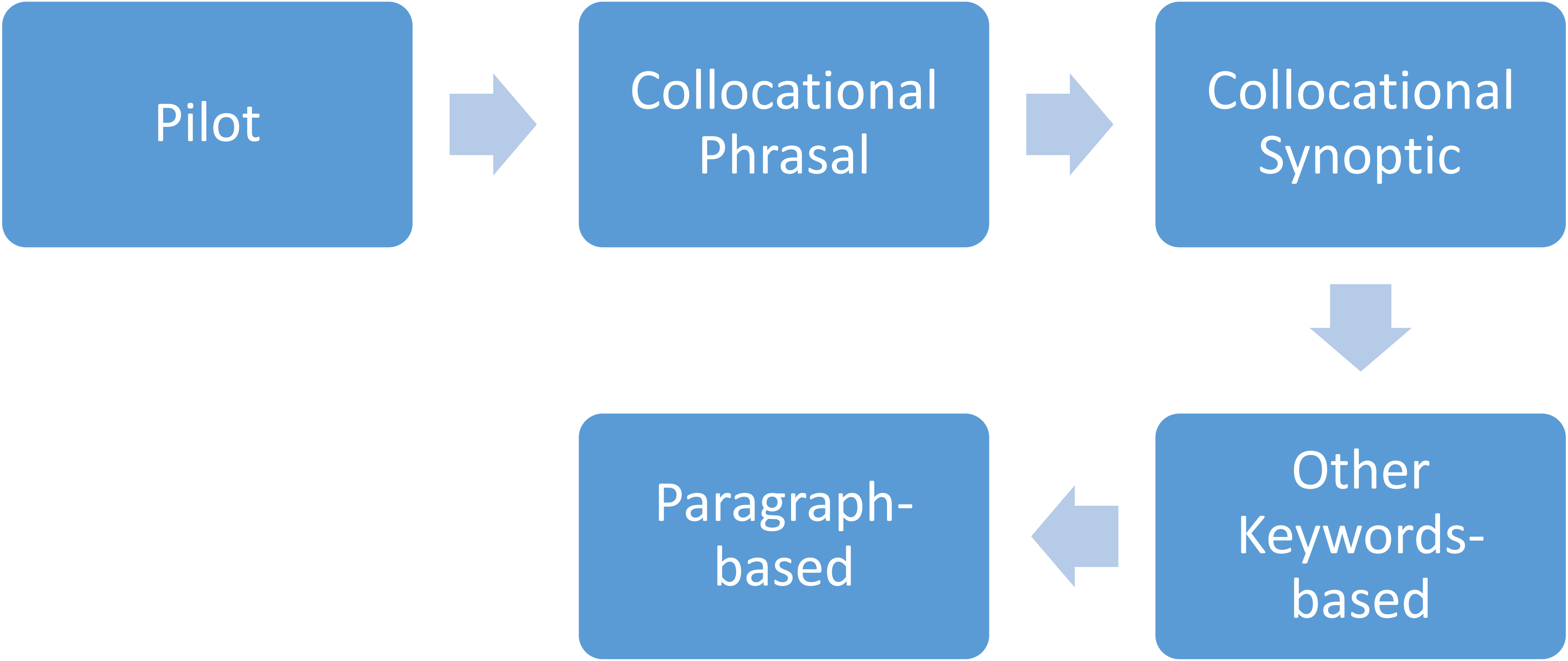
Keywords-based

- Pilot query
  - 6 best KW, ChatNoir, Indri
- Collocational Phrasal
  - 3 terms long collocations, Derived from the Pilot, Indri
- Collocational
  - Derived from the Pilot, 2 terms long collocations combined into 6 terms long queries, Chatnoir
- Other Keywords-based
  - Remaining KW, 6 terms long q., Chatnoir

Paragraph based

- Paragraph chunking
- One query from each paragraph
- Paragraph position [start, end], inside the document
- 10 terms with highest TF-IDF score from the whole paragraph
- Chatnoir

Queries Scheduling



Pilot

Collocational Phrasal

Collocational Synoptic

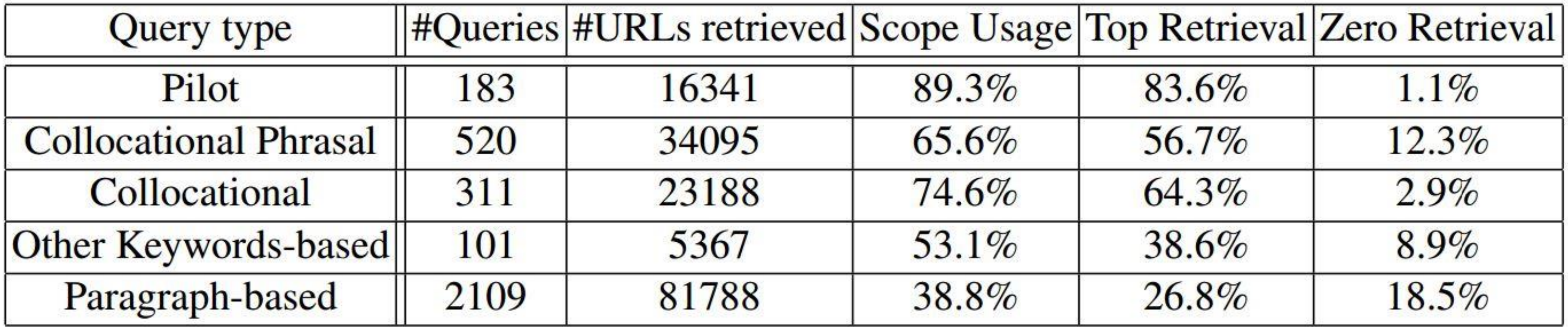
Other Keywords-based

Paragraph-based

Method Assessment During Test Phase

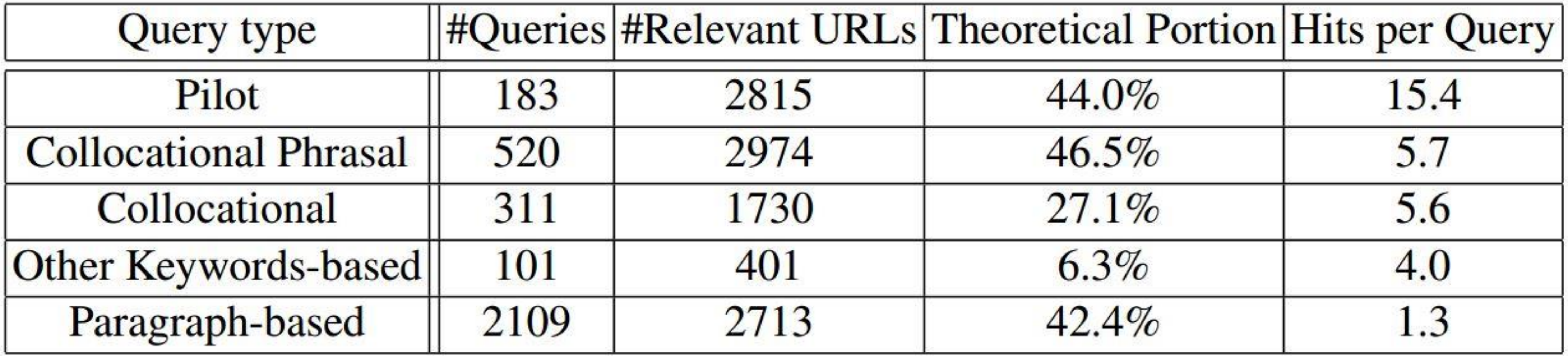
- 98 documents
- 32.9 queries per document on average
  - 18.8% directed to Indri, 81.2% to ChatNoir
- Max 100 URLs per one query
- 134 247 unique URLs retrieved in total
- 32 538 URLs downloaded
- 6 392 URLs were relevant
- Master hit as retrieval of an annotated URL
  - 0.45 recall, 5 documents with recall 1, and 12 documents with recall 0

Query Type Scope



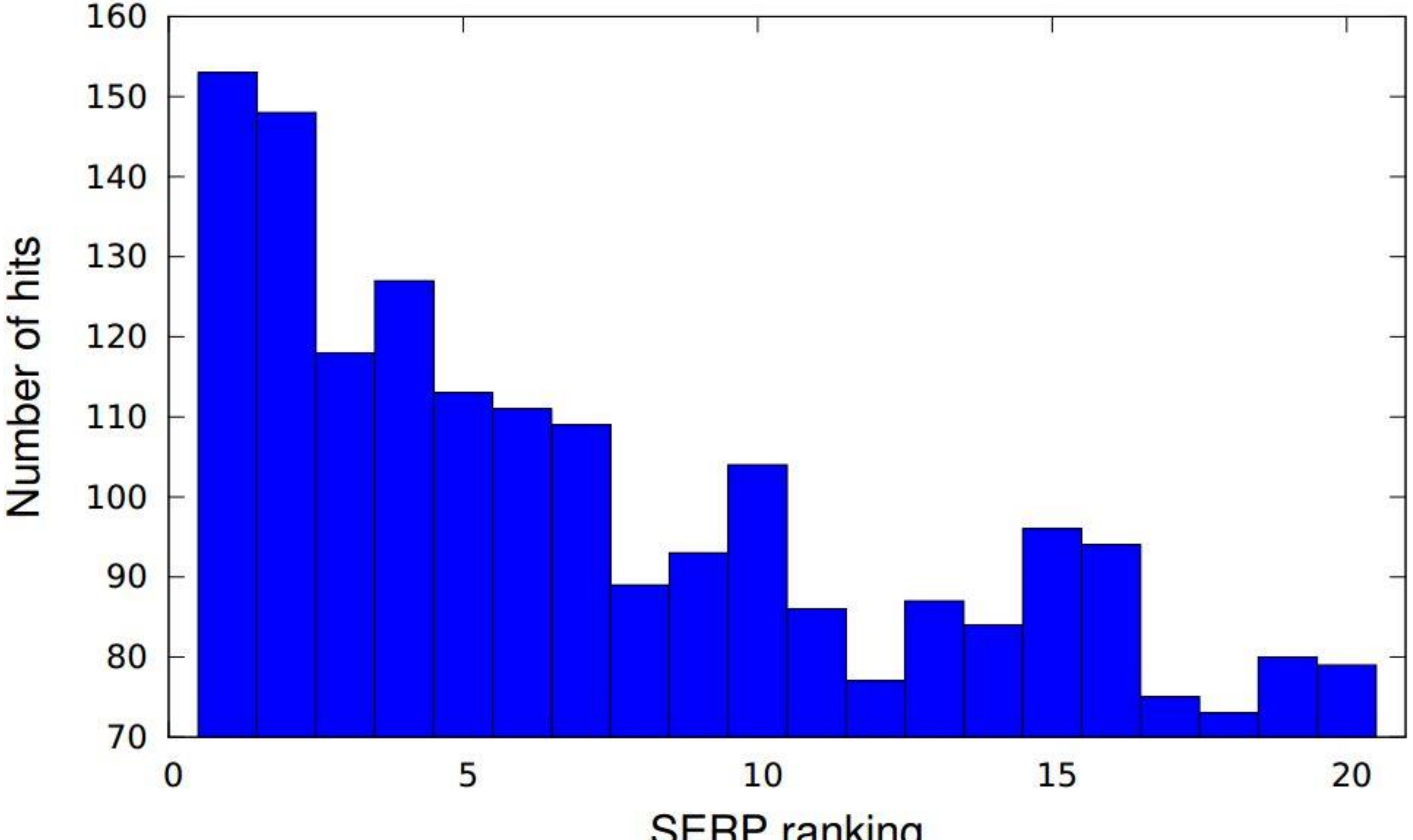
Query type	#Queries	#URLs retrieved	Scope Usage	Top Retrieval	Zero Retrieval
Pilot	183	16341	89.3%	83.6%	1.1%
Collocational Phrasal	520	34095	65.6%	56.7%	12.3%
Collocational	311	23188	74.6%	64.3%	2.9%
Other Keywords-based	101	5367	53.1%	38.6%	8.9%
Paragraph-based	2109	81788	38.8%	26.8%	18.5%

Query Type Performance

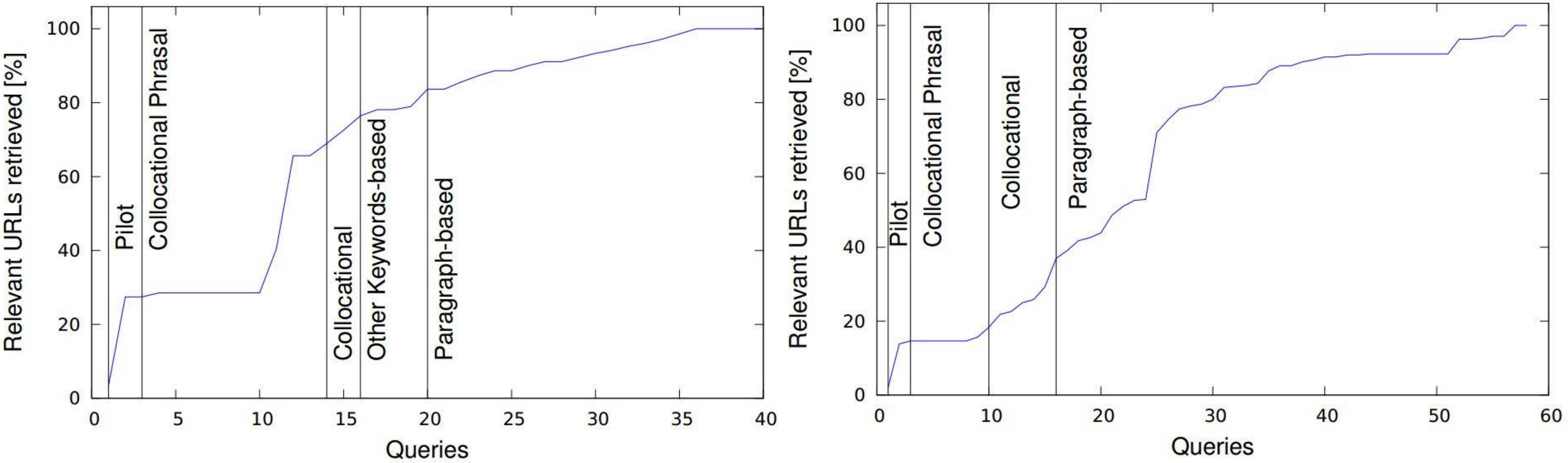


Query type	#Queries	#Relevant URLs	Theoretical Portion	Hits per Query
Pilot	183	2815	44.0%	15.4
Collocational Phrasal	520	2974	46.5%	5.7
Collocational	311	1730	27.1%	5.6
Other Keywords-based	101	401	6.3%	4.0
Paragraph-based	2109	2713	42.4%	1.3

Success Rate per SERP Rank



Source Retrieval Progress Based on 2 Selected Documents



Conclusions

- Usable methodology for source retrieval
- The pilot queries proved to be the best choice for synoptic search
- Paragraph-based queries perform well in position retrieval, but not well enough
- Achieved the highest recall among this year’s softwares