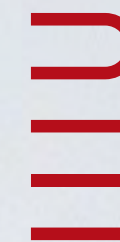


1222 • 2022
800
ANNI



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

What subject indexers can learn from CLEF

Nicola Ferro

Intelligent Interactive Information Access (IIA) Hub
Department of Information Engineering
University of Padua





Why Evaluation?



“To measure is to know”

“If you cannot measure it,
you cannot improve it”

Lord William Thompson,
first Baron Kelvin (1824-1907)

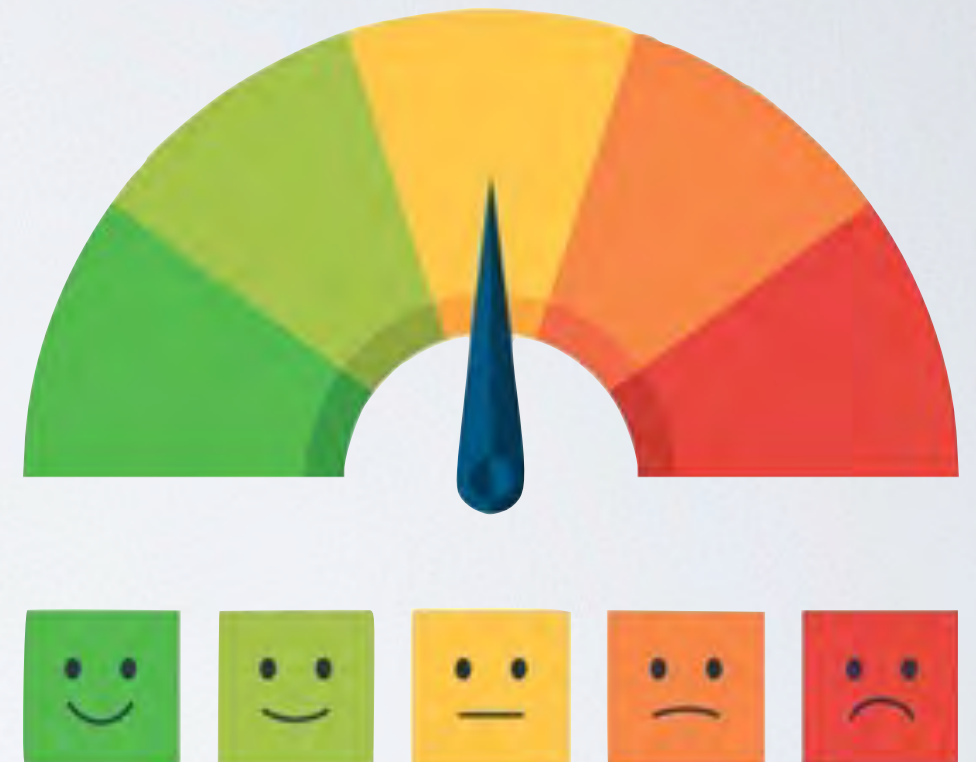
What to Evaluate?

Efficiency



VS

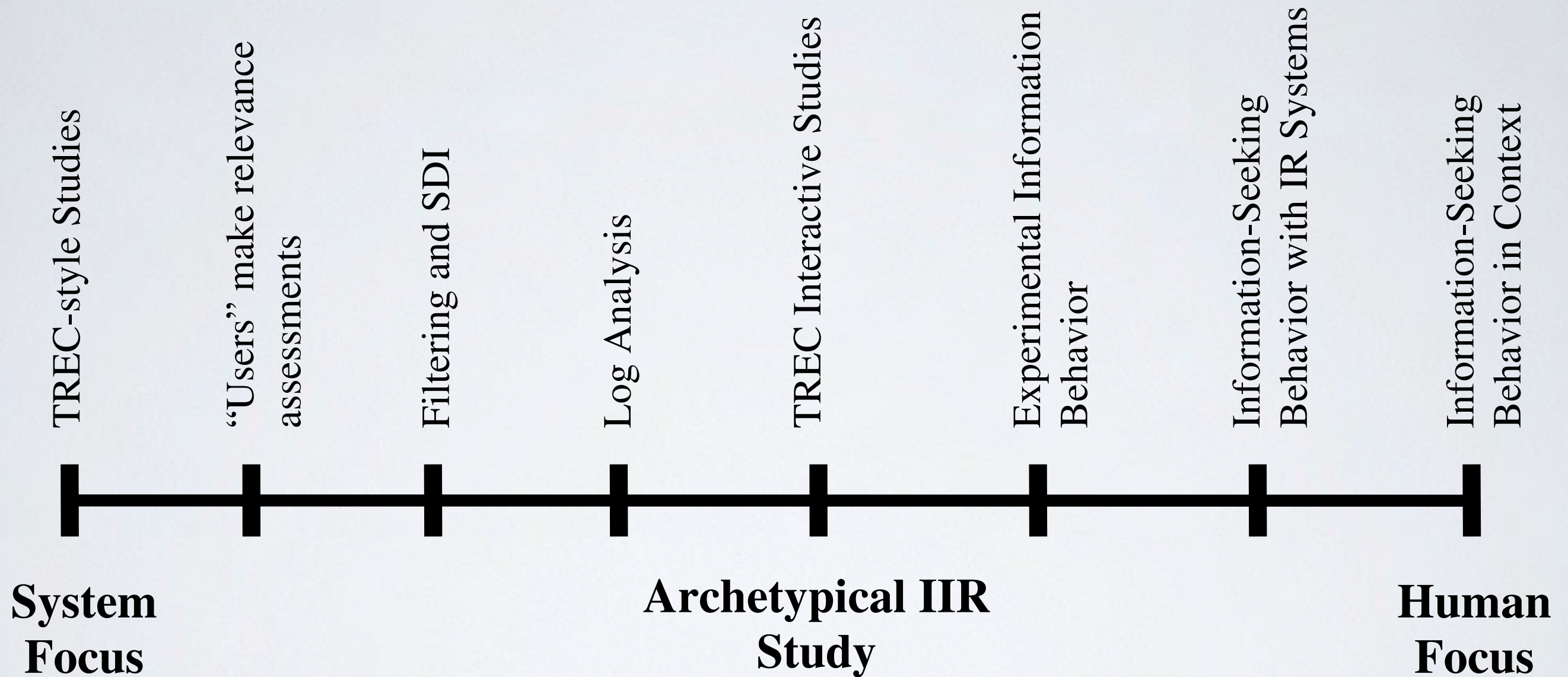
Effectiveness





- It must be scientifically **valid**
 - valid methodology, measures, and statistics
 - large-scale enough to be statistically valid
 - must be “repeatable” if possible
- It must be **realistic** for the applications that will be using the information retrieval systems
 - **task** and use cases
- It must be **understandable** to your audience/client

Harman, D. K. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA.



Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. Foundations and Trends in Information Retrieval (FnTIR), 3(1-2), 1-224.



The Beginning

- **Cranfield Paradigm** by Cyril W. Cleverdon

- Defines the use of **experimental collections**

- **documents** (corpora)
- **topics**, which are a surrogate for information needs
- **relevance judgments** (binary or graded) also called relevance assessment or ground-truth (or qrels)



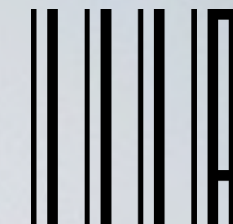
Cyril W. Cleverdon

- Ensures **comparability** and **reproducibility** of the experiments

Cleverdon, C. W. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Aslib Cranfield Research Project, College of Aeronautics, Cranfield, UK.

Cleverdon, C. W. (1997). *The Cranfield Tests on Index Languages Devices*. In Spärck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA.

The “Ideal Test Collection”



Corpora → (not historical) corpora are typically OK

- < 500 documents, no real value
- 1-2,000 documents, minimally acceptable
- > 10,000 documents, actually needed

Topics → typical size is still 50 topics

- < 75 topics, no real value
- 250 topics, minimally acceptable
- > 1,000 topics, actually needed

Relevance Judgements → binary is still most common option, diversity only recently

- multi-graded (highly and fairly relevant)
- types (novel, stimulating, ...)
- need for **pooling** (still open research issue)

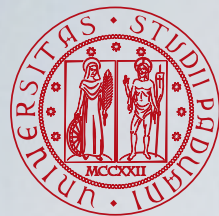


Karen Spärck Jones



C. J. “Keith” van Rijsbergen

Spärck Jones, K. and van Rijsbergen, C. J. (1975). *Report on the need for and provision of an ‘ideal’ information retrieval test collection*. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge.

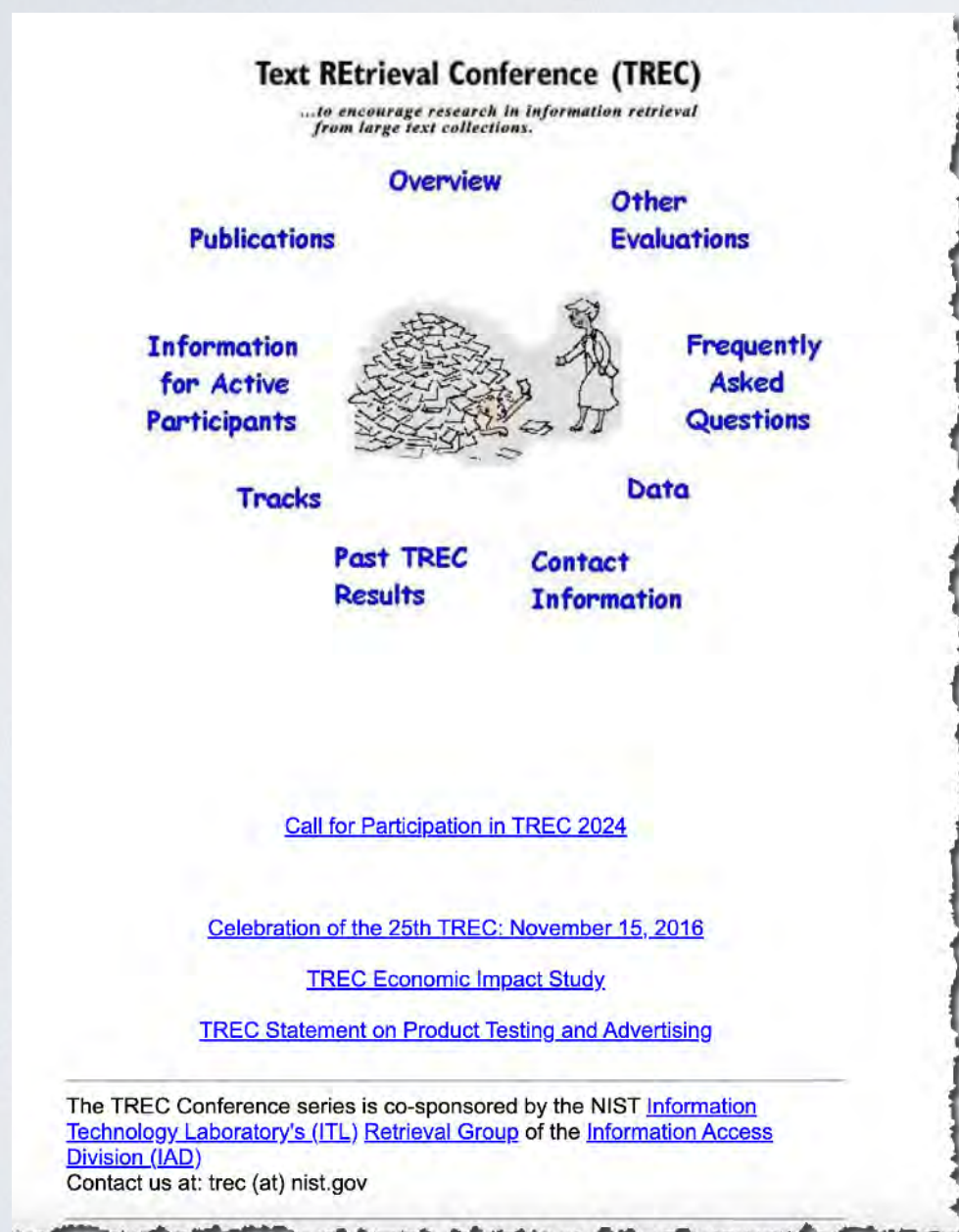


Large-scale Evaluation Initiatives: TREC



● **TREC** (Text REtrieval Conference), USA, since 1992

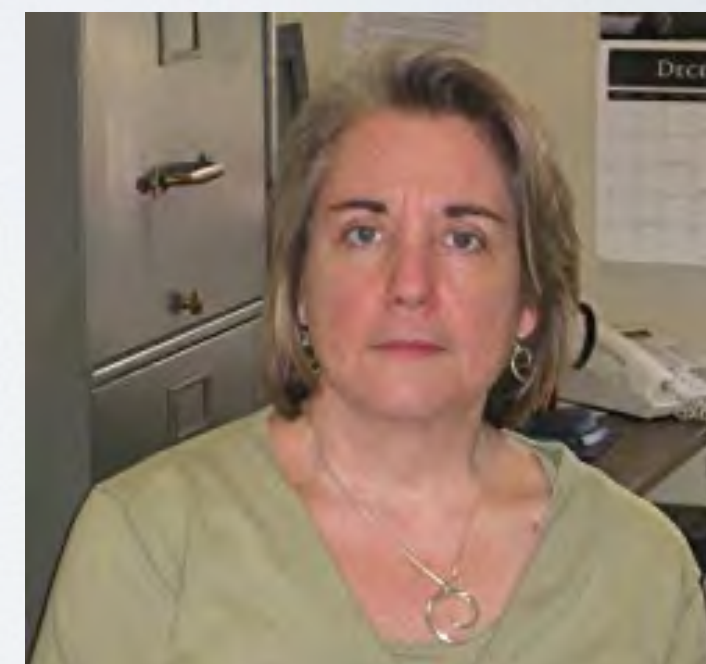
● <https://trec.nist.gov/>



Ian Soboroff

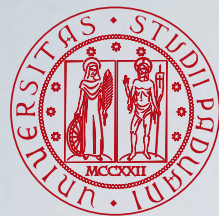


Donna Harman



Ellen M. Voorhees

Harman, D. K. and Voorhees, E. M., editors (2005). *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, USA.



Large-scale Evaluation Initiatives: NTCIR



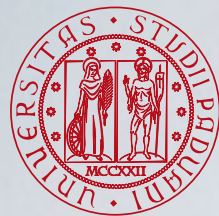
- **NTCIR** (NII Testbeds and Community for Information access Research), Japan, since 1999

● <http://research.nii.ac.jp/ntcir/index-en.html>



Noriko Kando

Sakai, T., Oard, D. W., and Kando, N., editors (2021). *Evaluating Information Retrieval and Access Tasks – NTCIR's Legacy of Research Impact*, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany.



Large-scale Evaluation Initiatives: CLEF

- **CLEF** (Conference and Labs of the Evaluation Forum), Europe, since 2000

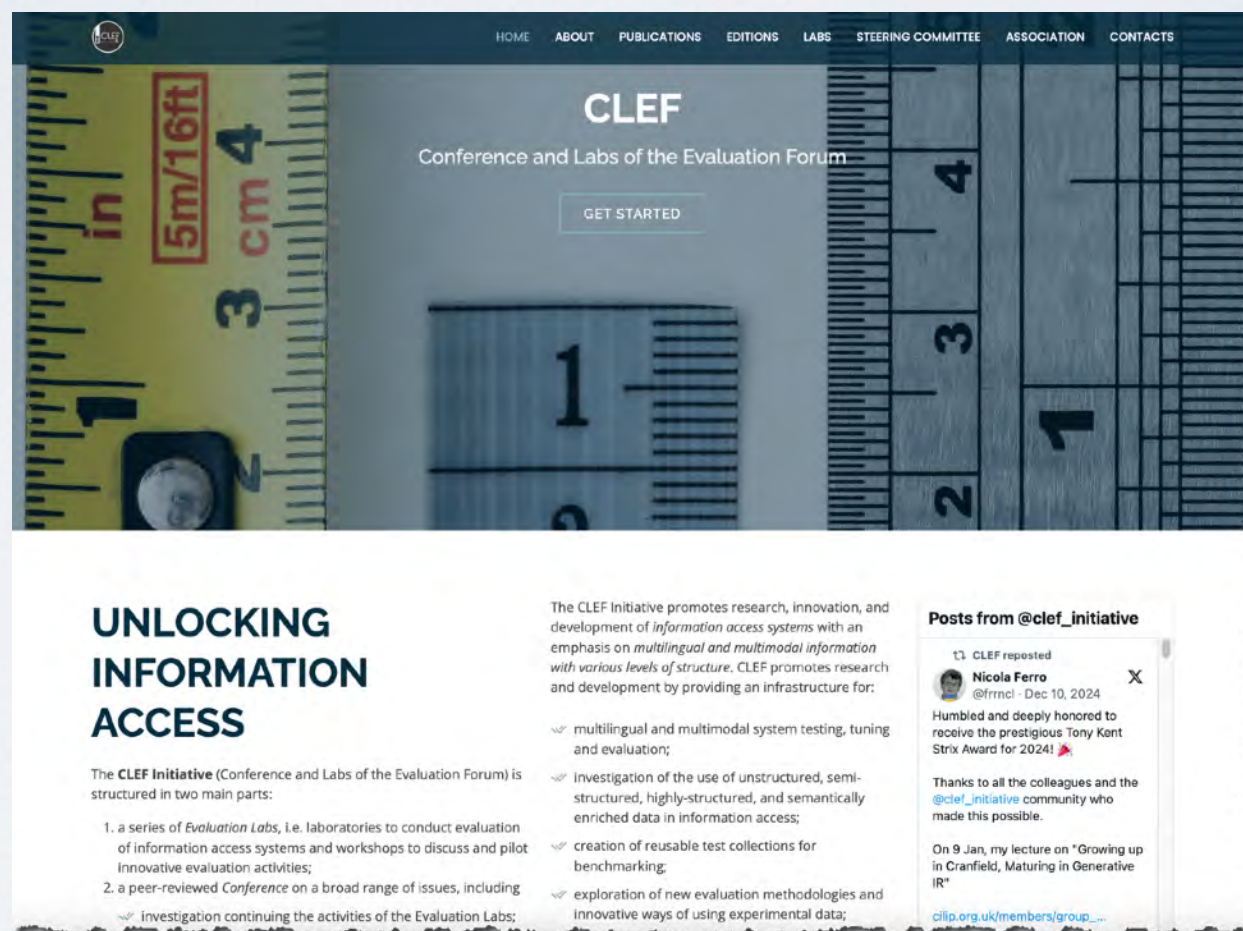
● <https://www.clef-initiative.eu/>



Alberto Barrón-Cedeño



Alba García Seco de Herrera

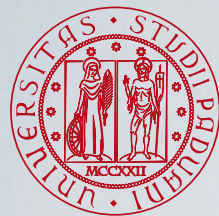


Carol Ann Peters



Nicola Ferro

Ferro, N. and Peters, C., editors (2019). *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany.



Large-scale Evaluation Initiatives: FIRE



- **FIRE** (Forum for Information Retrieval Evaluation), India, since 2008

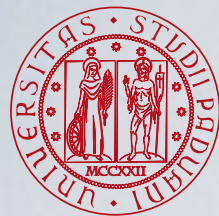
● <https://fire.irsirsi.org.in/>



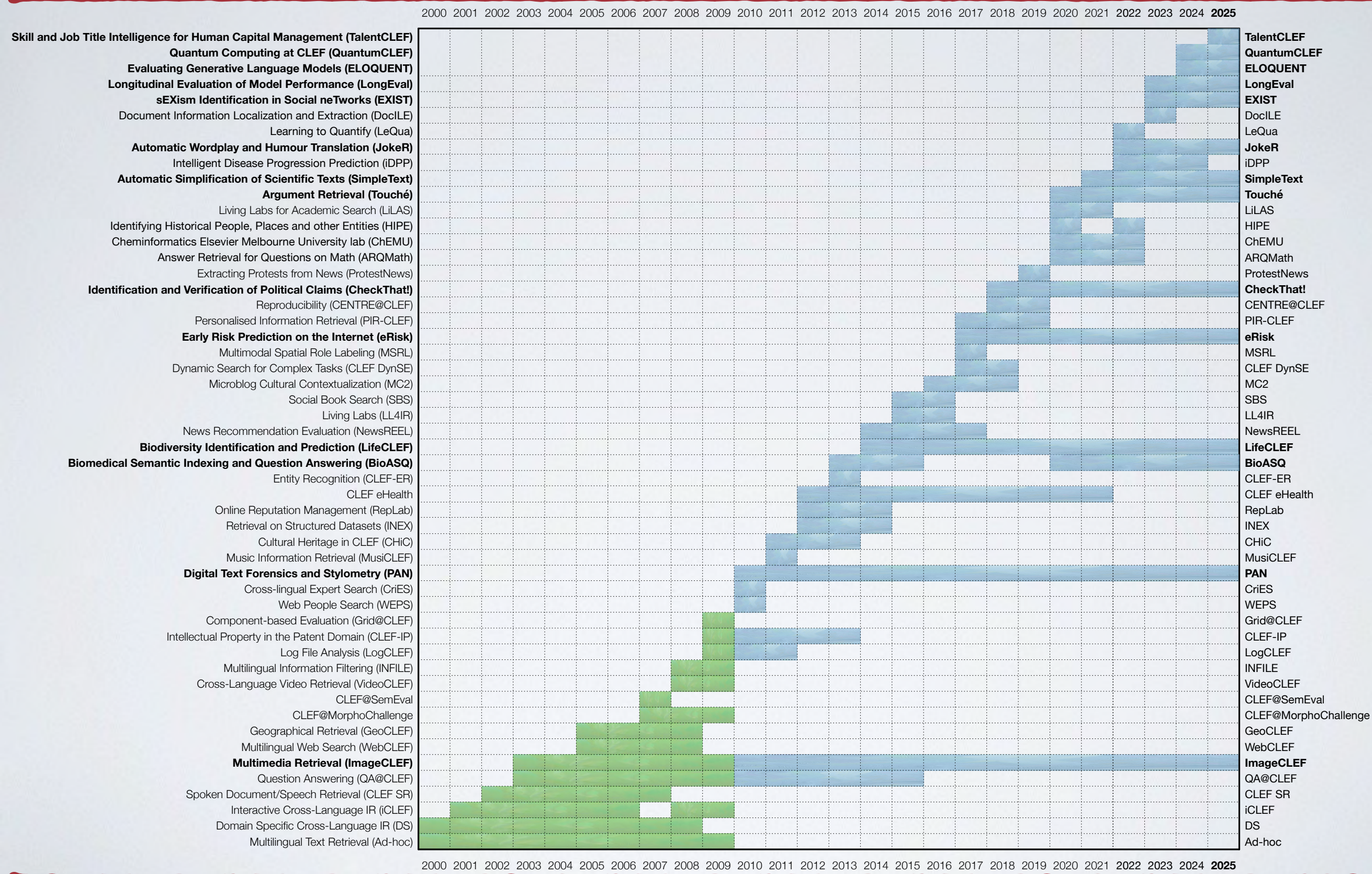
Mandar Mitra

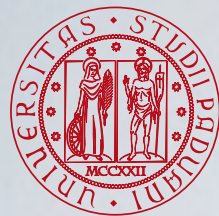


Prasenjit Majumder

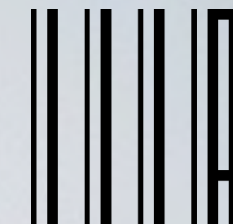


CLEF Labs over Time





Achievements: A Science of Ms



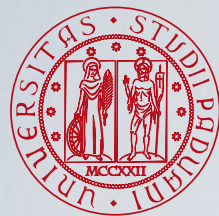
Methodologies

Models

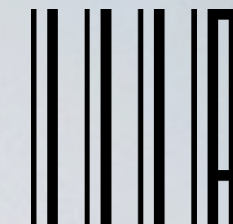
Measures

Major Systems

Massive Collections



How Valuable is Evaluation?



- The **TREC 2010 Economic Impact** study estimated in about **30 M\$** the overall **investment in TREC** by NIST
 - probably much much more if we had a means to estimate also the investment by participants in TREC
- They are the **pillars** for all the subsequent **scientific research** and **technology development**
 - TREC estimated the **return on investment** in the range of **3\$-5\$** for each invested dollar

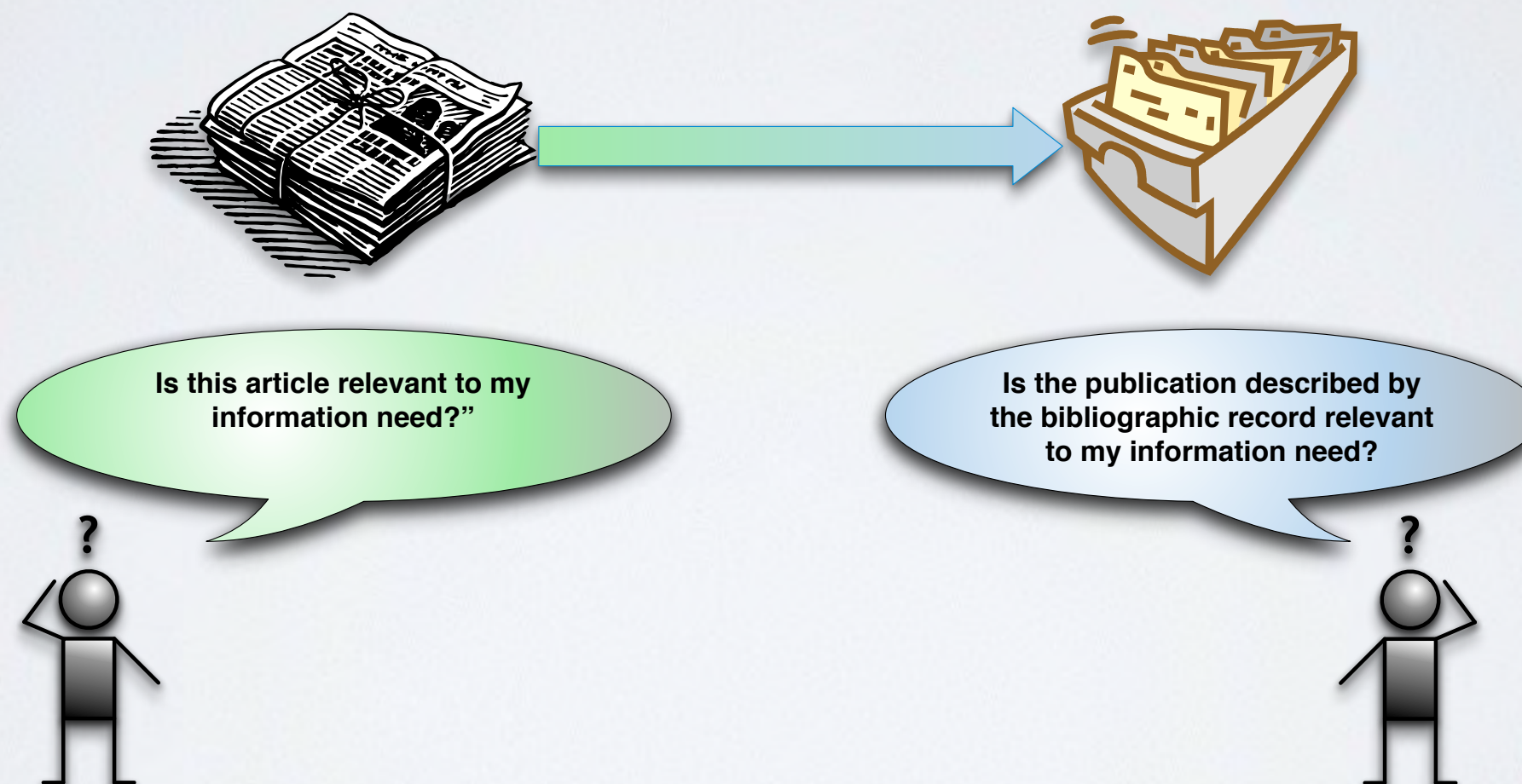


Rowe, B. R., Wood, D. W., Link, A. L., and Simoni, D. A. (2010). *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*. RTI Project Number 0211875, RTI International, USA. <http://trec.nist.gov/pubs/2010.economic.impact.pdf>.



CLEF 2008 and 2009: The European Library (TEL) Tasks

- The task is to search and retrieve relevant items from collections of **library catalog cards**, which are surrogates for documents held by libraries
- Both **monolingual** and **bilingual** tasks have been offered
- Not only the data are very sparse and less rich than newspapers but also the task is different from a traditional ad-hoc task

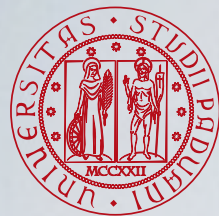


Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2009). CLEF 2008: Ad Hoc Track Overview. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., and Peñas, A., editors, *Evaluating Systems for Multilingual and Multimodal Information Access: Ninth Workshop of the Cross-Language Evaluation Forum (CLEF 2008). Revised Selected Papers*, pages 15–37. Lecture Notes in Computer Science (LNCS) 5706, Springer, Heidelberg, Germany.

Ferro, N. and Peters, C. (2010). CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In Peters, C., Di Nunzio, G. M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, 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 13–35. Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany.



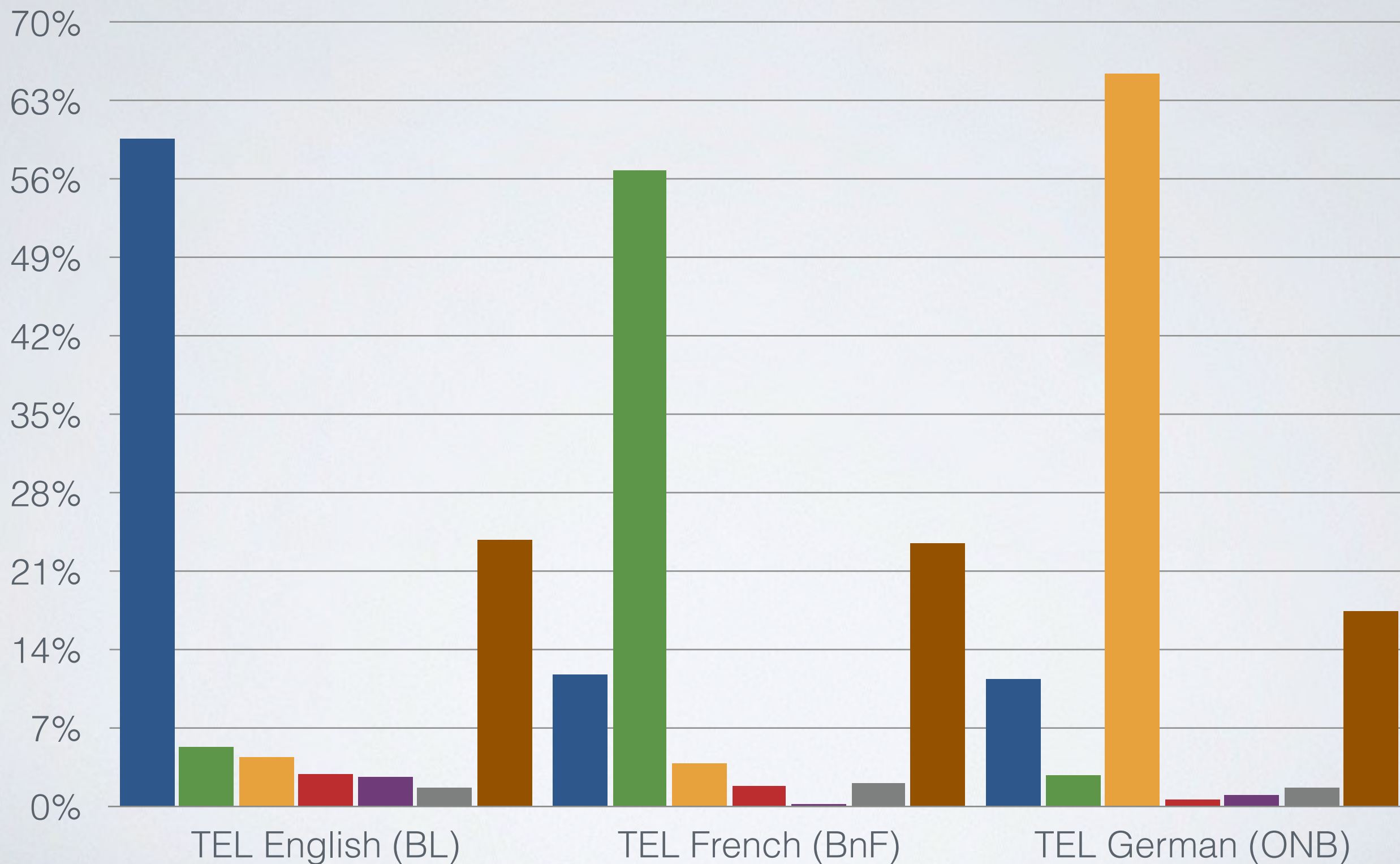
- The collections have been provided by The European Library (<http://www.theeuropeanlibrary.org/>) and are catalog records harvested from Europe's national libraries
- English
 - source: British Library (BL)
 - size: 1,208,383,351 bytes
 - items: 1,000,100 records
- French
 - source: Bibliothèque Nationale de France (BnF)
 - size: 1.362.122.091 bytes
 - items: 1,000,100 records
- German
 - source: Austrian National Library (ONB)
 - size: 1.306.492.248 bytes
 - items: 869,353 records

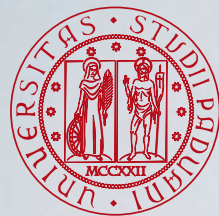


TEL Collections: Distribution of the Languages

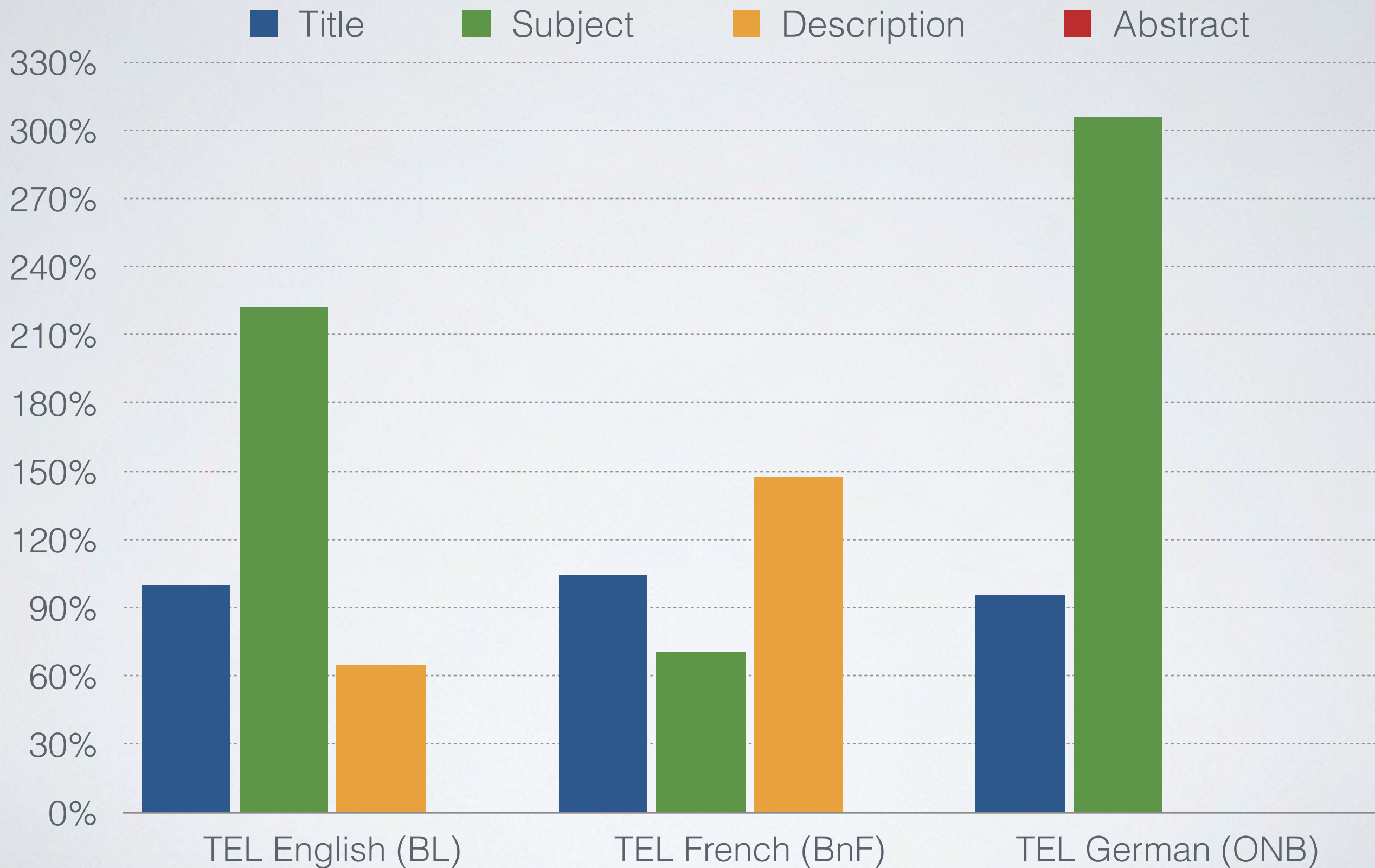


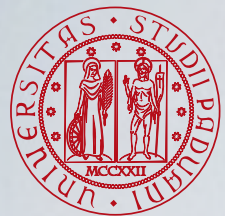
English French German Spanish Russian Italian Other





TEL Collections: Distribution of the Content

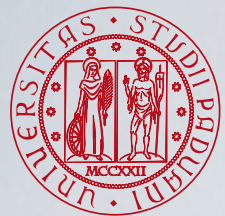




TEL Collections: English BL Sample Record

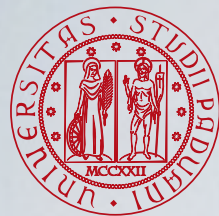


```
<record>
  <set>TEL_BL_opac</set>
  <header>
    <id>010734316</id>
  </header>
  <document format="index">
    <index>
      <topic>BL_opac</topic>
    </index>
  </document>
  <document format="dcx">
    <oai_dc:dc>
      <dc:title>Country wives </dc:title>
      <dc:contributor> Shaw, Rebecca, 1931- </dc:contributor>
      <dc:publisher>London : Orion, c2001</dc:publisher>
      <dc:terms:issued>c2001</dc:terms:issued>
      <dc:terms:extent>263 p. : 1 map ; 20 cm.</dc:terms:extent>
      <dc:language xsi:type="IS0639-2">eng</dc:language>
      <dc:description>Originally published: 2001.</dc:description>
      <dc:abstract>Country Wives is the second in a series of novels set in a veterinary practice on the outskirts of a Dorset town, by the bestselling author of the Turnham Malpas novels.</dc:abstract>
      <dc:subject xsi:type="dcterms:DDC">823.914</dc:subject>
      <dc:subject>Veterinarians</dc:subject>
      <dc:identifier xsi:type="lib:ISBN">0752844733 (pbk.) :</dc:identifier>
      <dc:type>text</dc:type>
      <dc:identifier
        xlink:href="http://catalogue.bl.uk/F/-?func=direct-doc-set&amp;l_base=BLL01&amp;from=TELgateway&amp;doc_number=010734316"
        >010734316</dc:identifier>
      <dc:identifier xsi:type="dcterms:URI"
        >http://catalogue.bl.uk/F/-?func=direct-doc-set&amp;l_base=BLL01&amp;from=TELgateway&amp;doc_number=010734316</dc:identifier>
      <mods:location>British Library HMNTS H.2002/3337</mods:location>
    </oai_dc:dc>
  </document>
</record>
```



TEL Collections: French BnF Sample Record

```
<record>
  <set>TEL_BnF_opac</set>
  <id>oai:bnf.fr:catalogue/ark:/12148/cb346737691/description</id>
  <document format="index">
    <index>
      <topic>BnF_opac</topic>
    </index>
  </document>
  <document format="dcx">
    <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:dc="http://purl.org/dc/elements/1.1/">
      <dc:identifiant>http://catalogue.bnf.fr/ark:/12148/cb346737691/description</dc:identifiant>
      <dc:title>Chine : la terre, l'eau et les hommes / écrit par Han Suyin ; photographié par
        Claude Sauvageot</dc:title>
      <dc:creator>Han, Suyin (1917-....)</dc:creator>
      <dc:contributor>Sauvageot, Claude (1935-....). Illustrateur</dc:contributor>
      <dc:publisher>Édition J.A. (Paris)</dc:publisher>
      <dc:date>1980</dc:date>
      <dc:description>Collection : Grands livres</dc:description>
      <dc:subject xml:lang="fre"> Chine -- Histoire</dc:subject>
      <dc:subject xml:lang="fre"> Chine -- Ouvrages illustrés</dc:subject>
      <dc:description>ISBN 2852581973</dc:description>
      <dc:language>fre</dc:language>
      <dc:type xml:lang="fre">texte imprimé</dc:type>
      <dc:type xml:lang="eng">printed text</dc:type>
      <dc:type xml:lang="eng">text</dc:type>
      <dc:rights xml:lang="fre">Catalogue en ligne de la Bibliothèque nationale de France</dc:rights>
      <dc:rights xml:lang="eng">French National Library online Catalog</dc:rights>
    </oai_dc:dc>
  </document>
</record>
```



TEL Collections: German ONB Sample Record



```
<record>
  <set>TEL_ONB_onb01</set>
  <id>oai:aleph.onb.ac.at:ONB01-000628794</id>
  <document format="index">
    <index>
      <topic>ONB_onb01</topic>
    </index>
  </document>
  <document format="dcx">
    <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:dcterms="http://purl.org/dc/dc/terms/"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:identifier xsi:type="onb:ACRecordId">AC03970200</dc:identifier>
      <dc:language xsi:type="dcterms:ISO639-2">ger</dc:language>
      <dc:creator>Salema, Teresa</dc:creator>
      <dc:title>&lt;&lt;Des&gt;&gt; Widerspenstigen Zähmung in der
        Gesellschaft Wilhelm Meisters: Ordnung der Natur oder Ironie der Kultur</dc:title>
      <dcterms:issued>1986</dcterms:issued>
      <dc:identifier xsi:type="onb:CallNumber">834301-B.NEU-Per.C,31</dc:identifier>
      <dcterms:bibliographicCitation>&lt;&lt;Der&gt;&gt; Widerspenstigen
        Zähmung</dcterms:bibliographicCitation>
      <dc:publisher xsi:type="onb:PlaceofPublisher">Innsbruck</dc:publisher>
      <dcterms:issued>1986</dcterms:issued>
      <dcterms:bibliographicCitation>(1986), S.143 - 156</dcterms:bibliographicCitation>
      <dc:subject>Romantik</dc:subject>
      <dc:subject>Feministische Literaturwissenschaft</dc:subject>
      <dc:subject>Klassik</dc:subject>
      <dc:subject>Frauenbild</dc:subject>
      <dc:subject>Weiblichkeitsideal</dc:subject>
      <dc:subject>Rollenbild</dc:subject>
      <dc:subject>Goethe, Johann Wolfgang &lt;&lt;von&gt;&gt;; Wilhelm
        Meisters Lehr- und Wanderjahre</dc:subject>
      <dcterms:abstract>Die feministische Literaturwissenschaft der letzten Jahre hat sich
        erfolgreich bemüht, die patriarchalisch orientierten Denk- und Handlungsstrategien
        aufzudecken, die seit der Aufklärung der Befreiung des Subjekts als Citoyen und
        Bourgeois zugrundeliegen. Daß eine solche "Bfreiung" sowohl mit der Negierung der
        Frau als Subjekt als auch mit der Unterdrückung von für weiblichgehaltenen Werten im
        Mann zusammenhängt, ist bisher und nicht nur aus feministischer Perspektive
        weitgehend bestätigt worden. Daher erscheint es sinnvoll, nach der Notwendigkeit
        einer solchen "Zähmung" grundsätzlicher zu fragen, indem Alternativmodelle aus jener
        Zeit analysiert werden, die zwar einen Zivilisationsprozeß mit dazugehöriger
        Triebbändigung postulieren, dies jedoch nicht aus eindeutig sexistischer Sicht tun.
        Goethes "Wilhelm Meister" ist hier in doppelter Hinsicht ein positives
        Gegenbeispiel: seiner frauenfreundlichen Einstellung wegen, und auch weil die beiden
        Romane im Gegensatz zur Literatur der Romantik mit der Wirklichkeit in positiver
        Konfrontation stehen</dcterms:abstract>
    </oai_dc:dc>
  </document>
</record>
```



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<topic>
  <identifier>10.2452/451-AH</identifier>

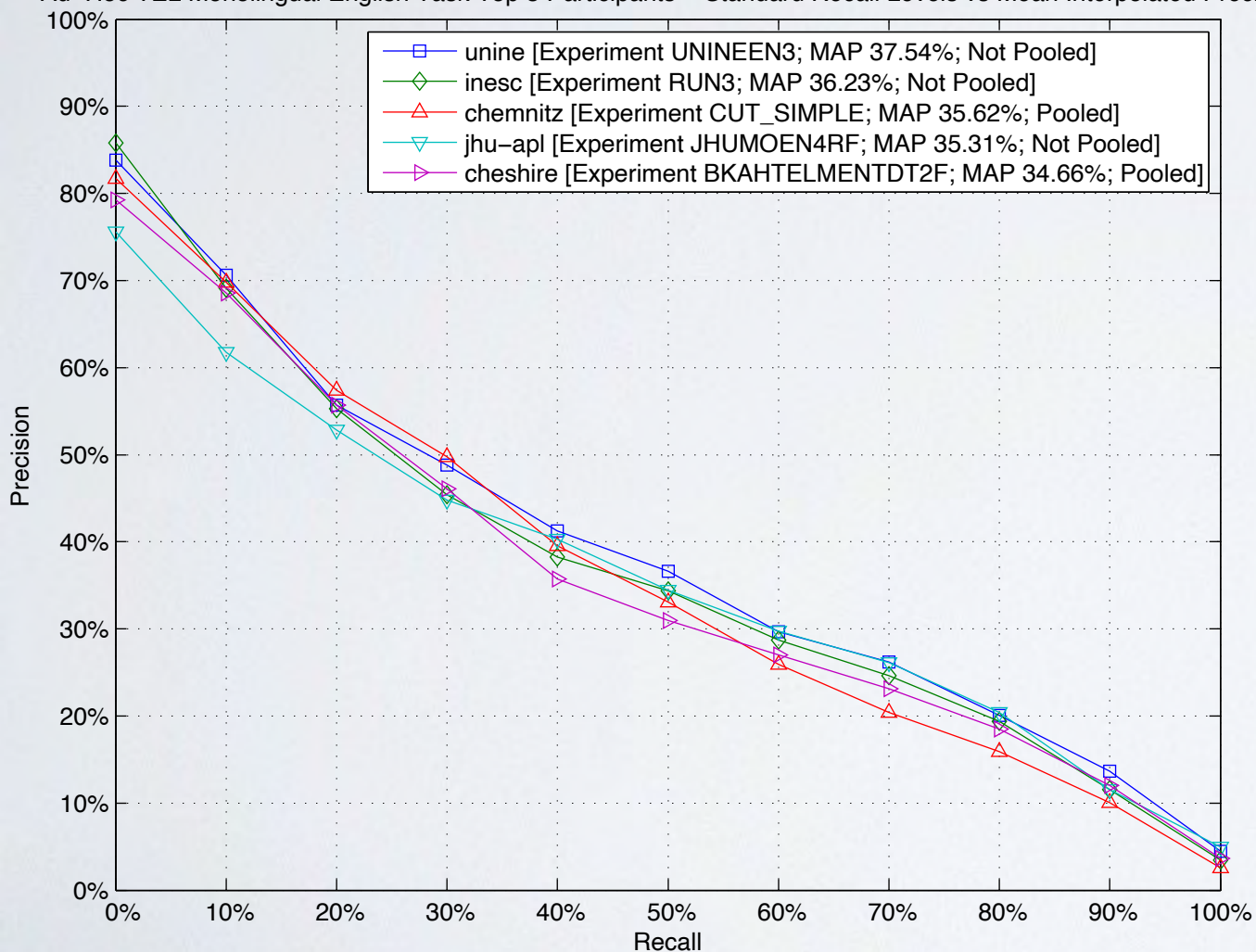
  <title lang="en">Roman Military in Britain</title>
  <title lang="de">Römisches Militär in Britannien</title>
  <title lang="es">El ejército romano en Britania</title>
  <title lang="fr">L'armée romaine en Grande-Bretagne</title>
  <title lang="nl">Romeinse Leger in Groot-Brittannie</title>

  <description lang="en">Find books or publications on the Roman invasion or military occupation
    of Britain.</description>
  <description lang="de">Finden Sie Bücher oder Publikationen über die römische Invasion oder das
    Militär in Britannien.</description>
  <description lang="es">Encuentre libros o publicaciones sobre la invasión romana o la ocupación
    militar romana en Britania.</description>
  <description lang="fr">Trouver des livres ou des publications sur l'invasion et l'occupation de
    la Grande-Bretagne par les Romains.</description>
  <description lang="nl">Vind boeken of publicaties over de Romeinse invasie of bezetting van
    Groot-Brittannie.</description>
</topic>
```

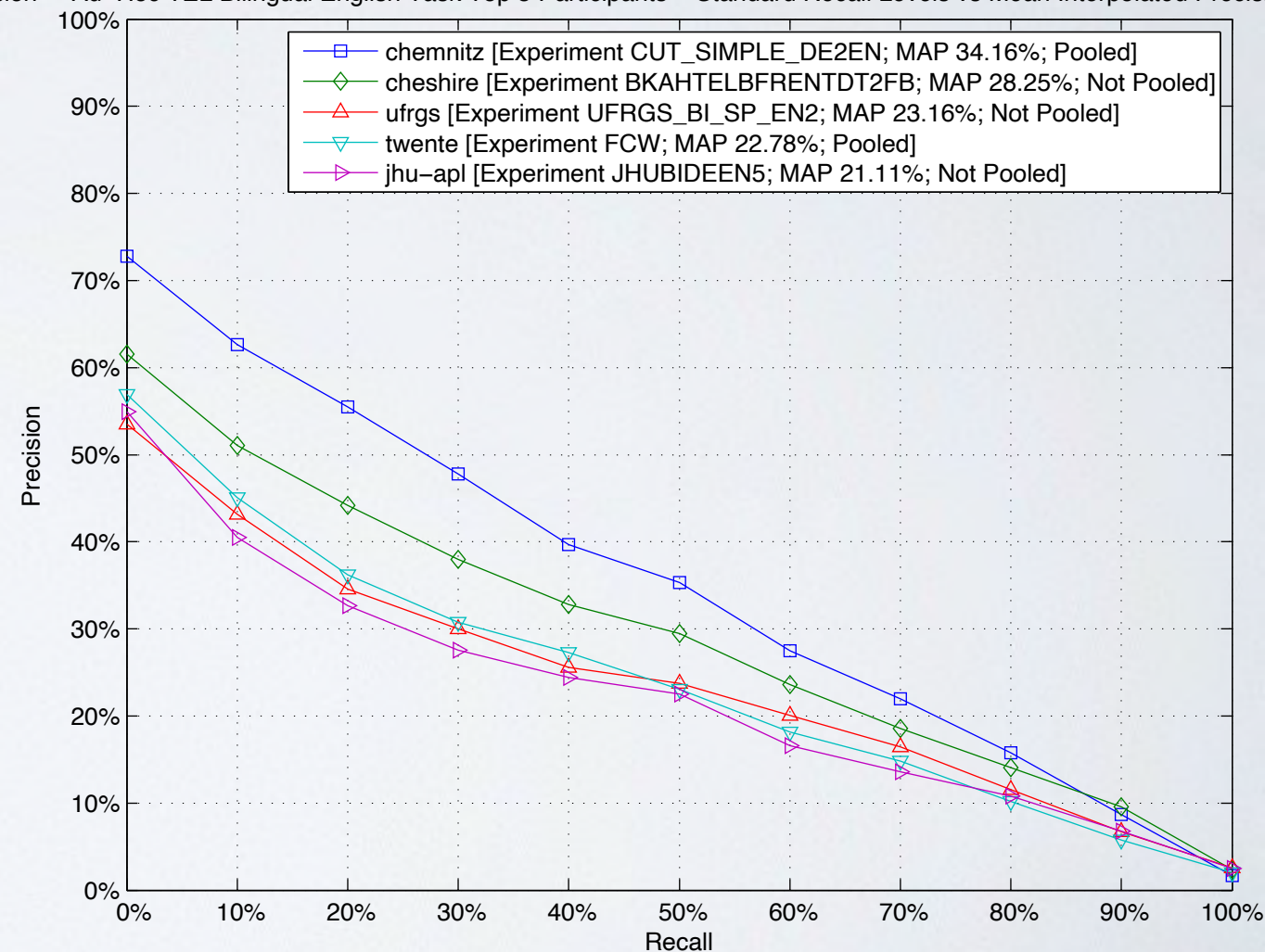
- 50 topics have been developed in English, German, and French both in CLEF 2008 and 2009
- Additional translations to Dutch and Spanish have been provided upon request
- Topics consist of **title** and **description** only; the **narrative** contained information relevant only to assessors

Bilingual is
91%
of Monolingual

Ad-Hoc TEL Monolingual English Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

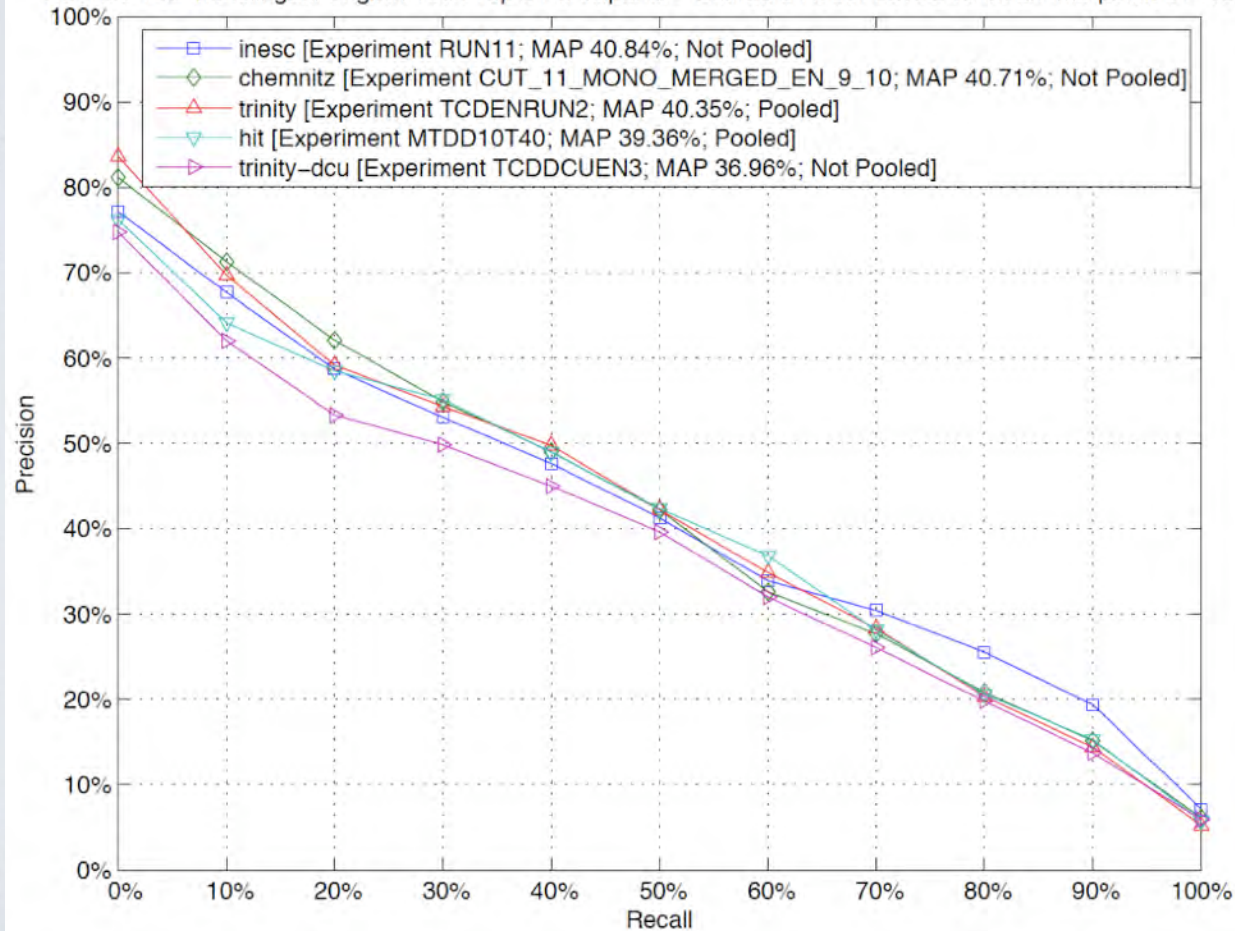


Ad-Hoc TEL Bilingual English Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

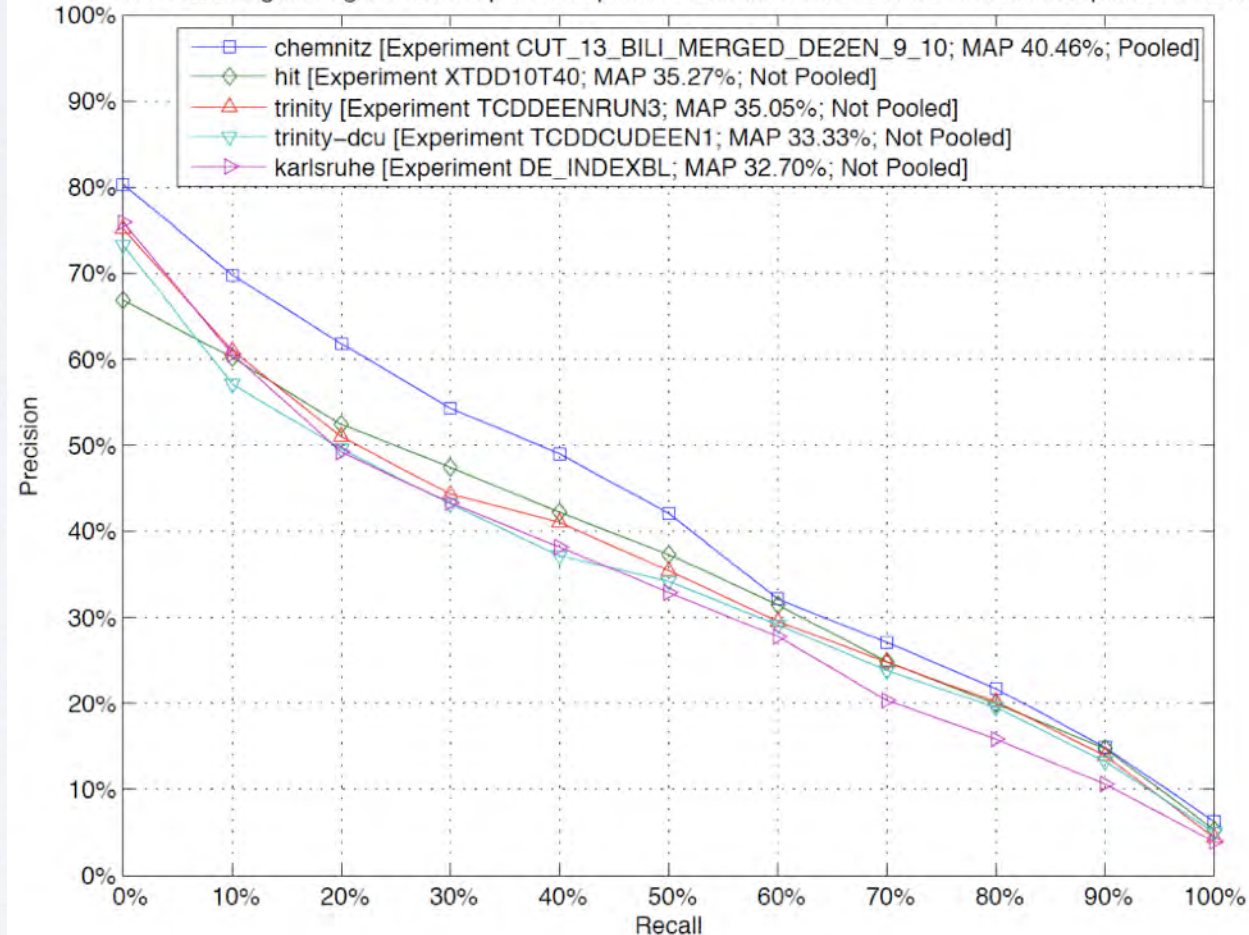


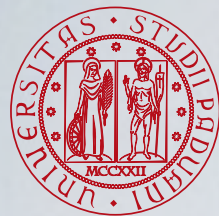
Bilingual is
99%
(vs 91% in 2008)
of Monolingual

Ad-Hoc TEL Monolingual English Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision



Ad-Hoc TEL Bilingual English Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision



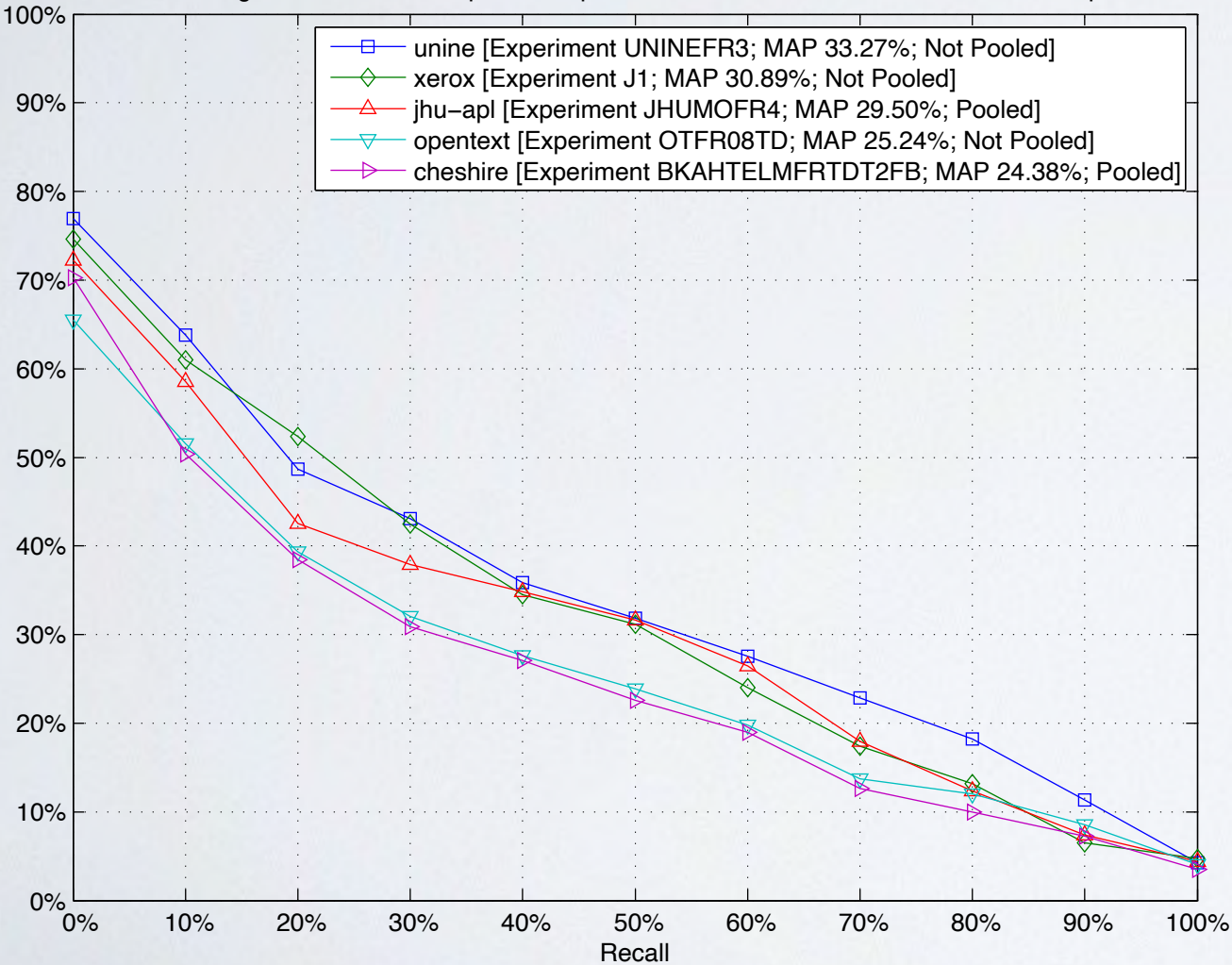


TEL 2008 French

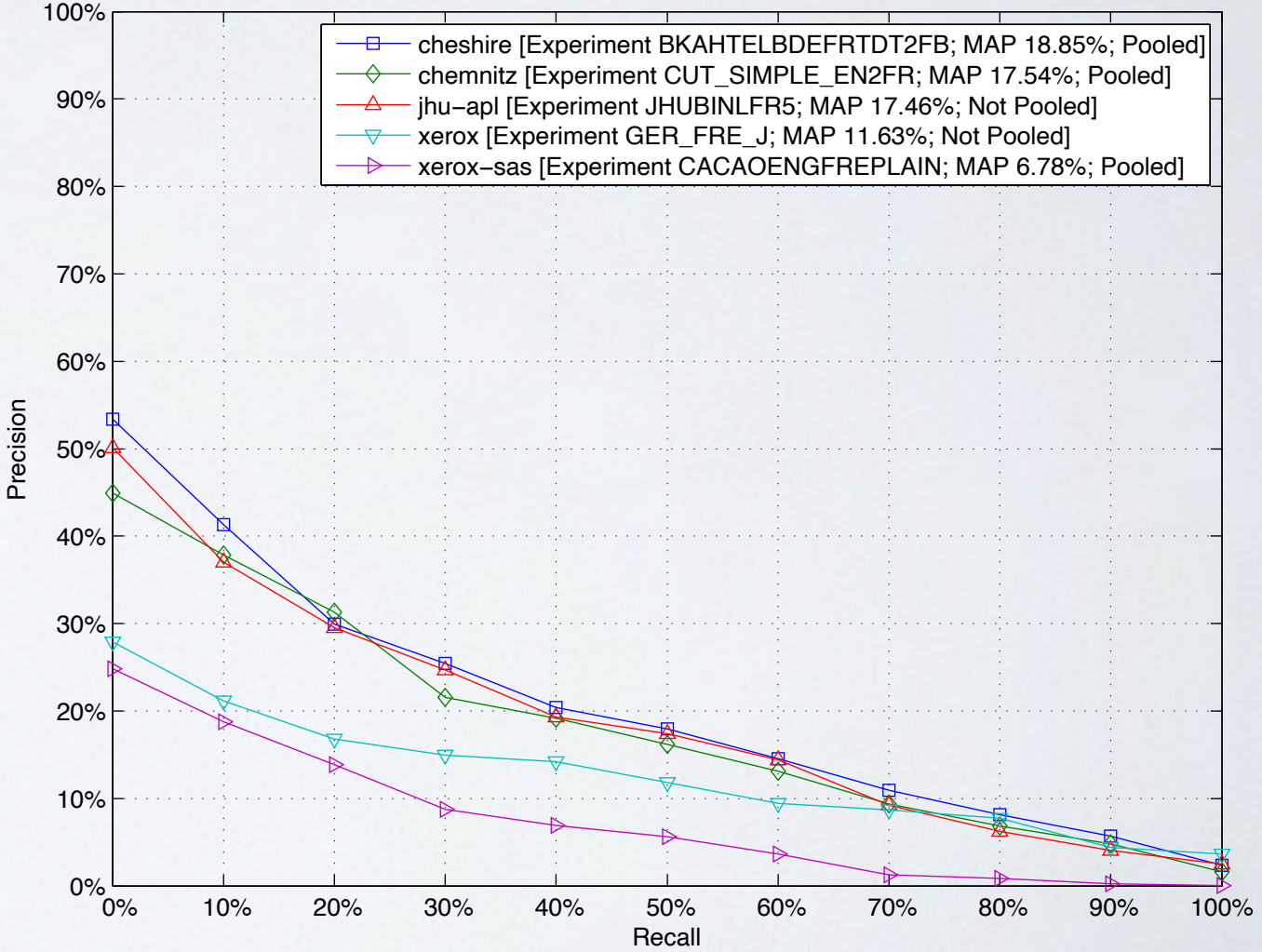


Bilingual is
57%
of Monolingual

Ad-Hoc TEL Monolingual French Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

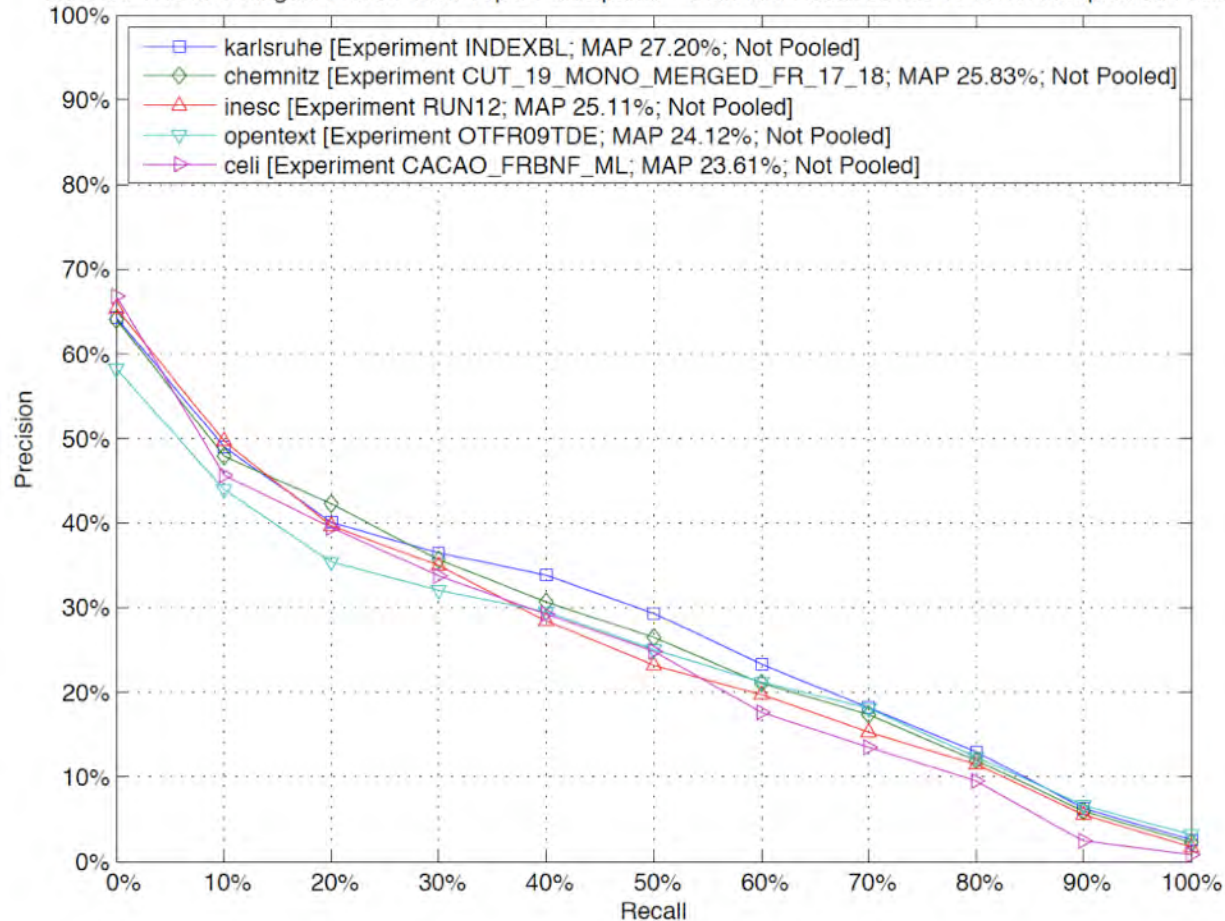


Ad-Hoc TEL Bilingual French Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

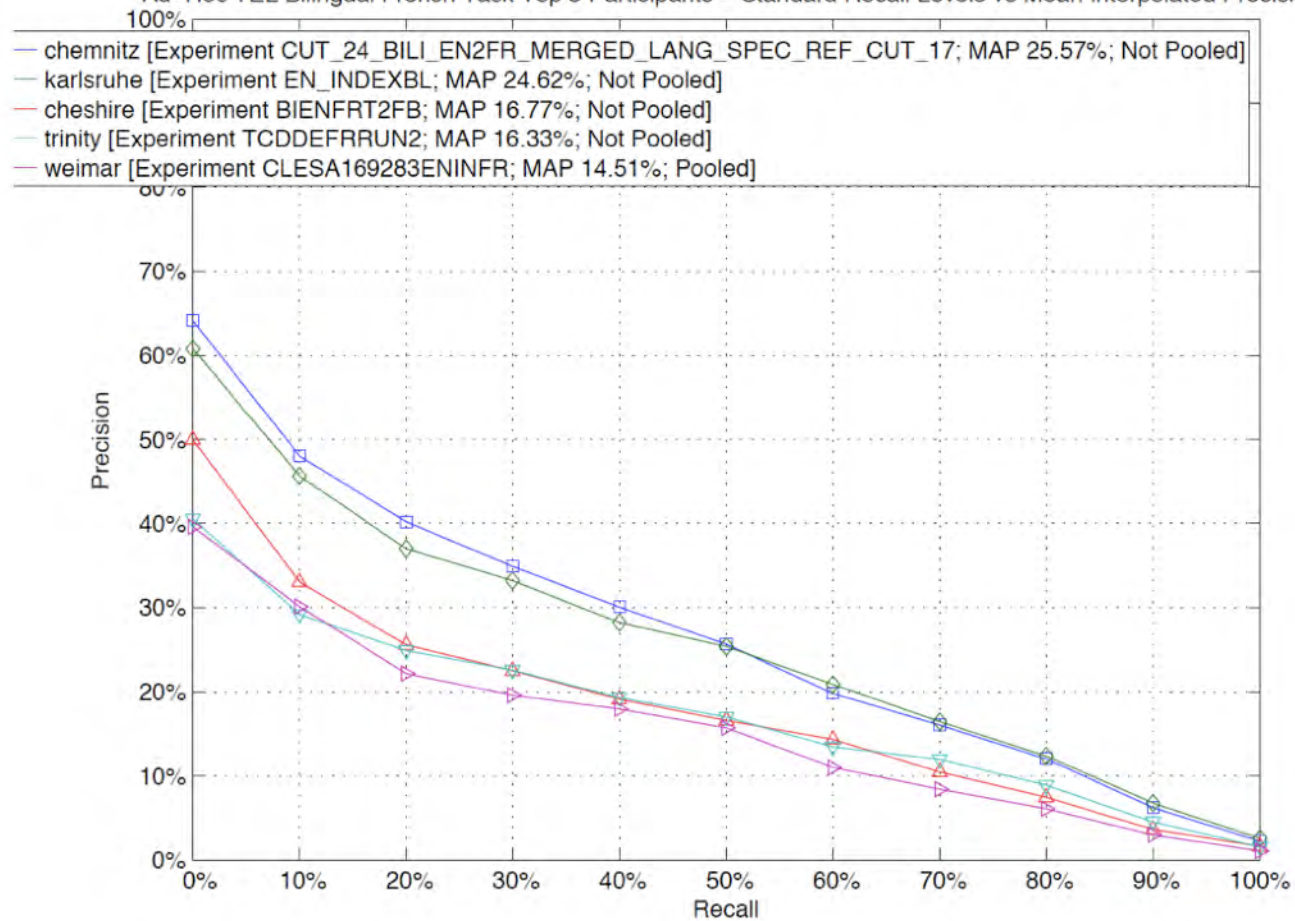


Bilingual is
94%
(vs 57% in 2008)
of Monolingual

Ad-Hoc TEL Monolingual French Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

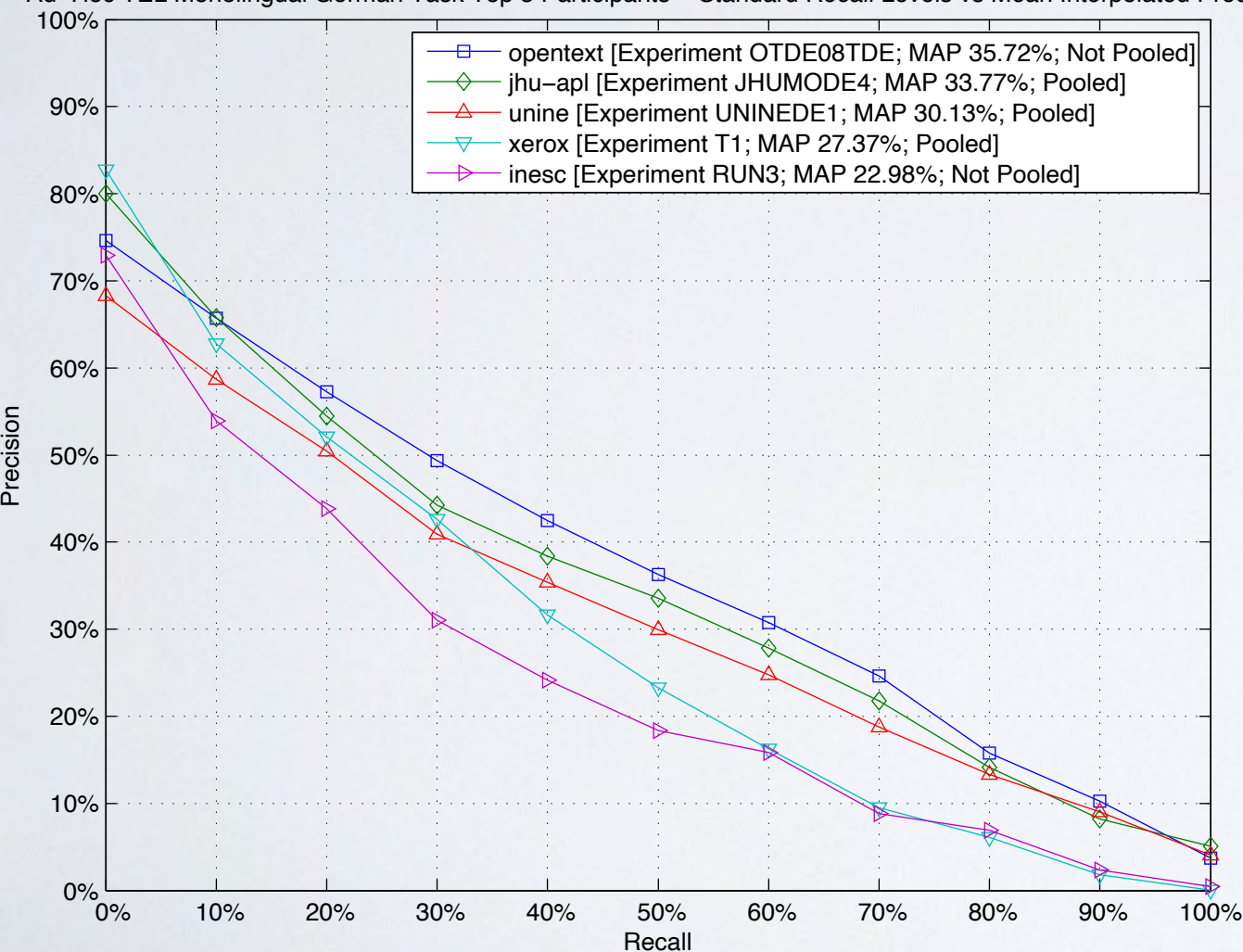


Ad-Hoc TEL Bilingual French Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision

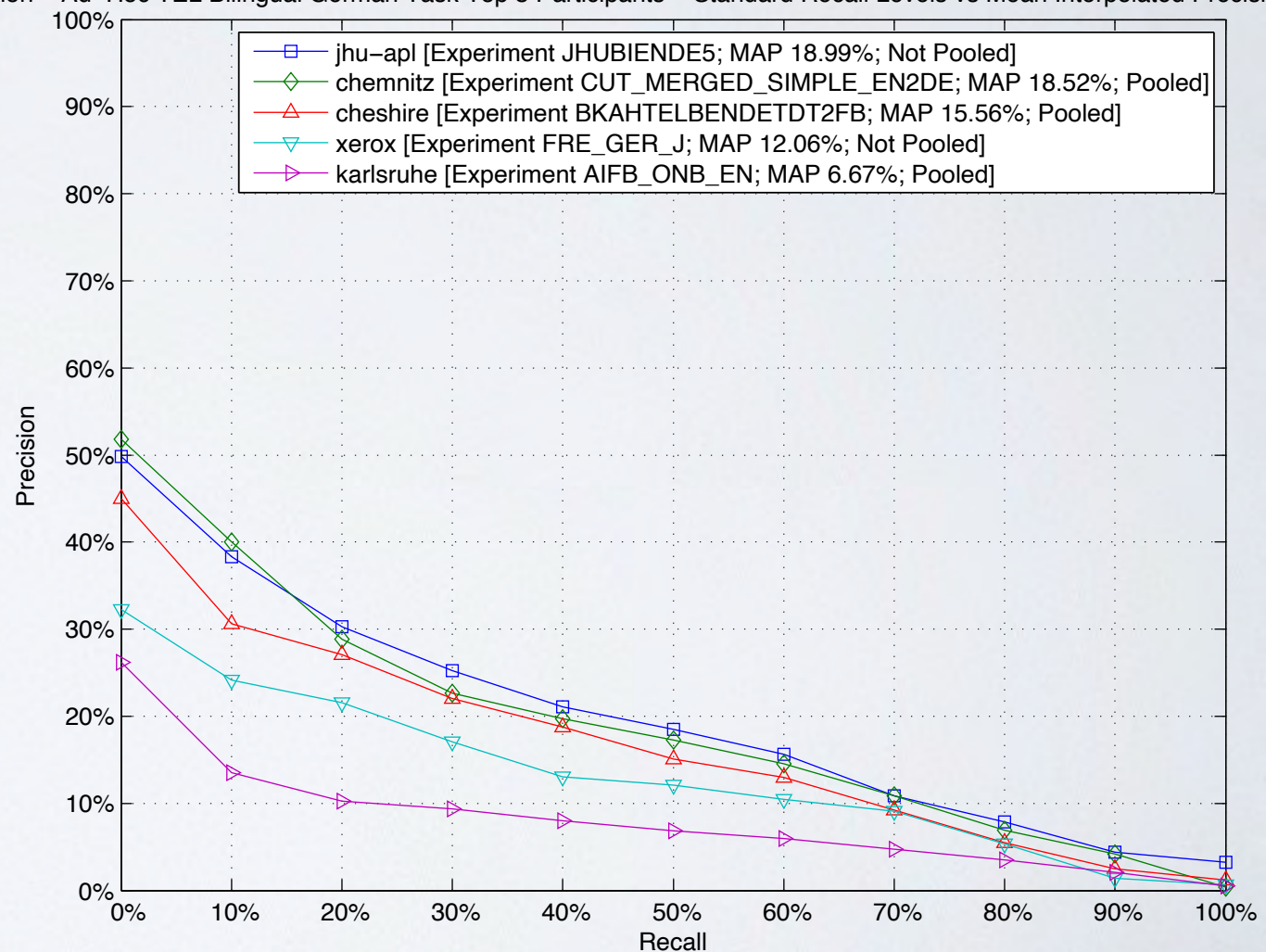


Bilingual is
53%
of Monolingual

Ad-Hoc TEL Monolingual German Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision



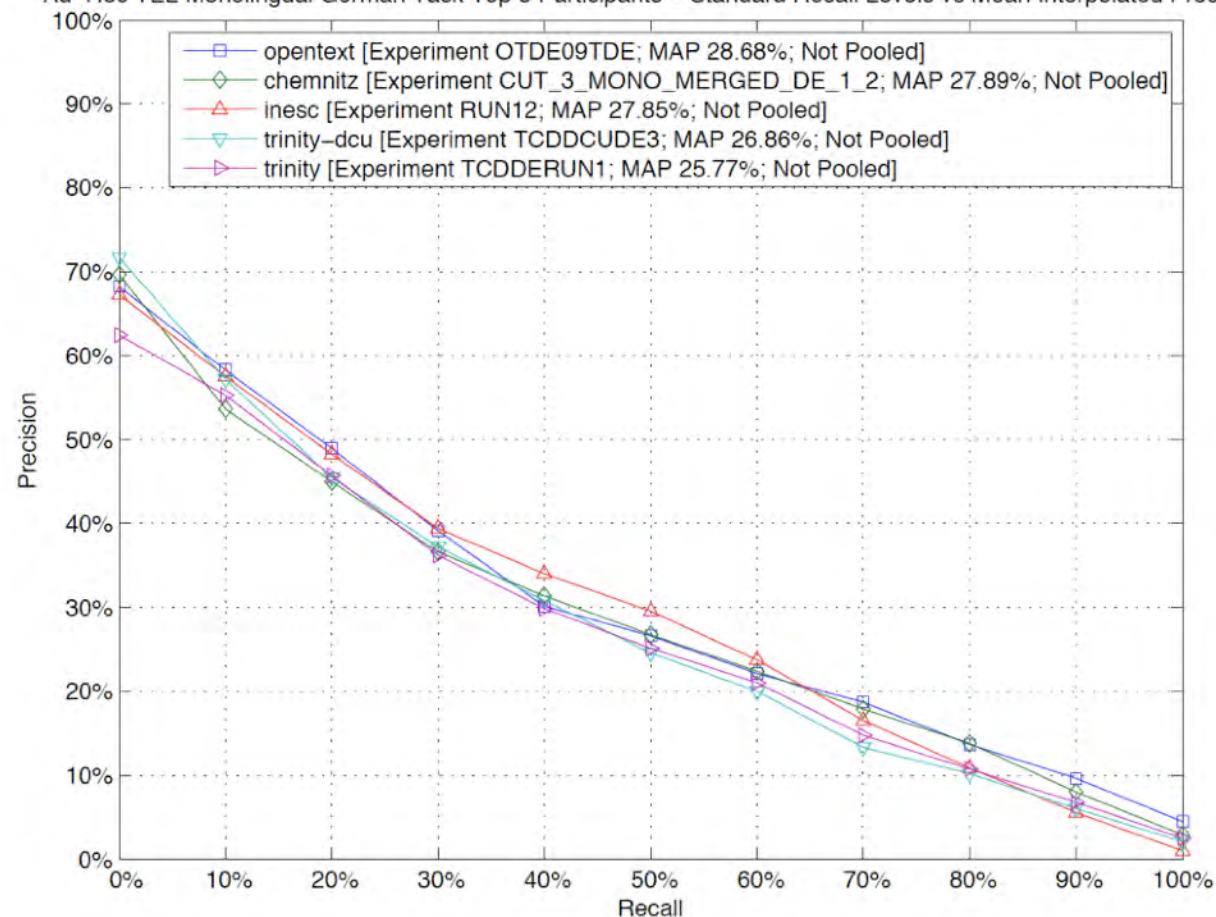
Ad-Hoc TEL Bilingual German Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision



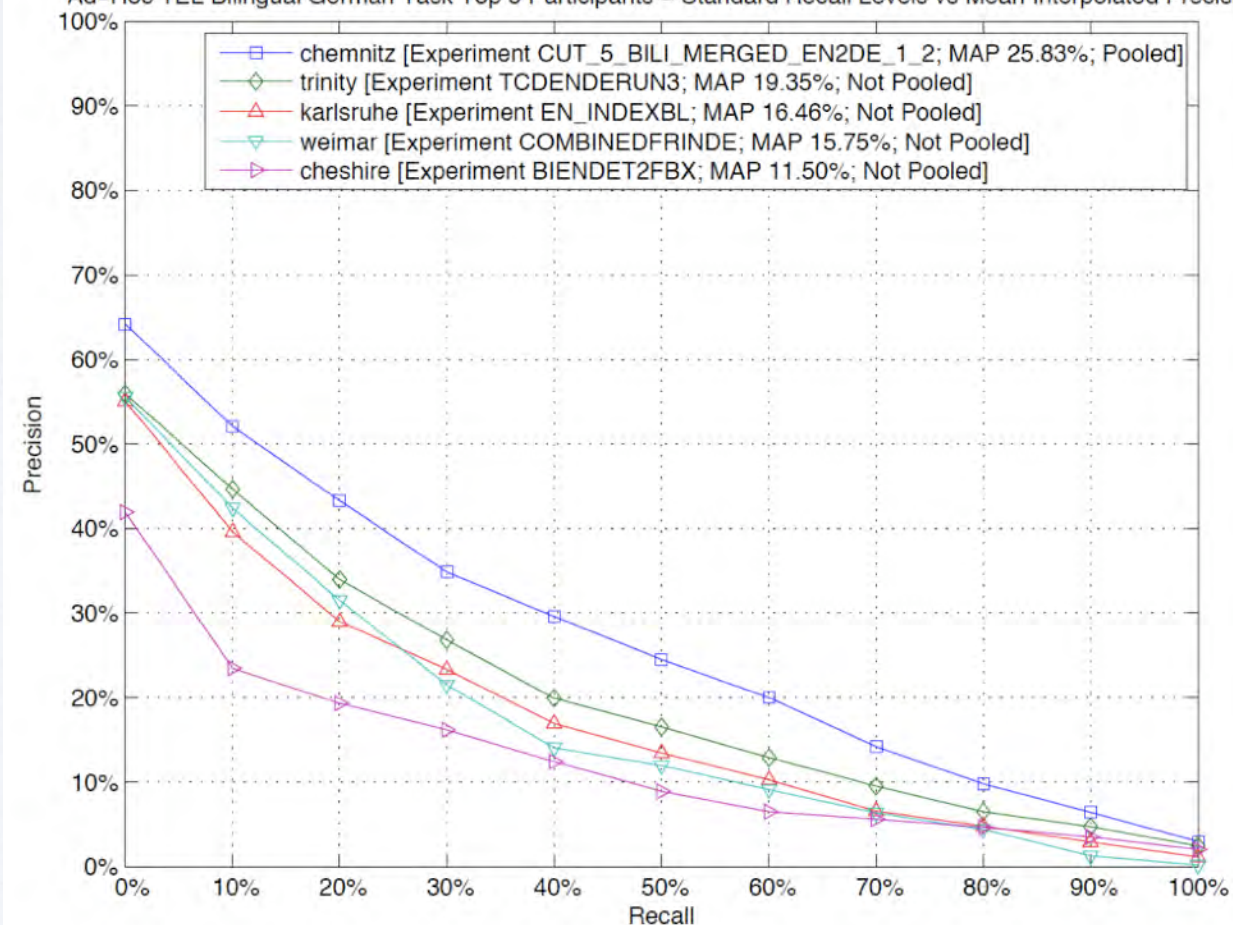
TEL 2009 German

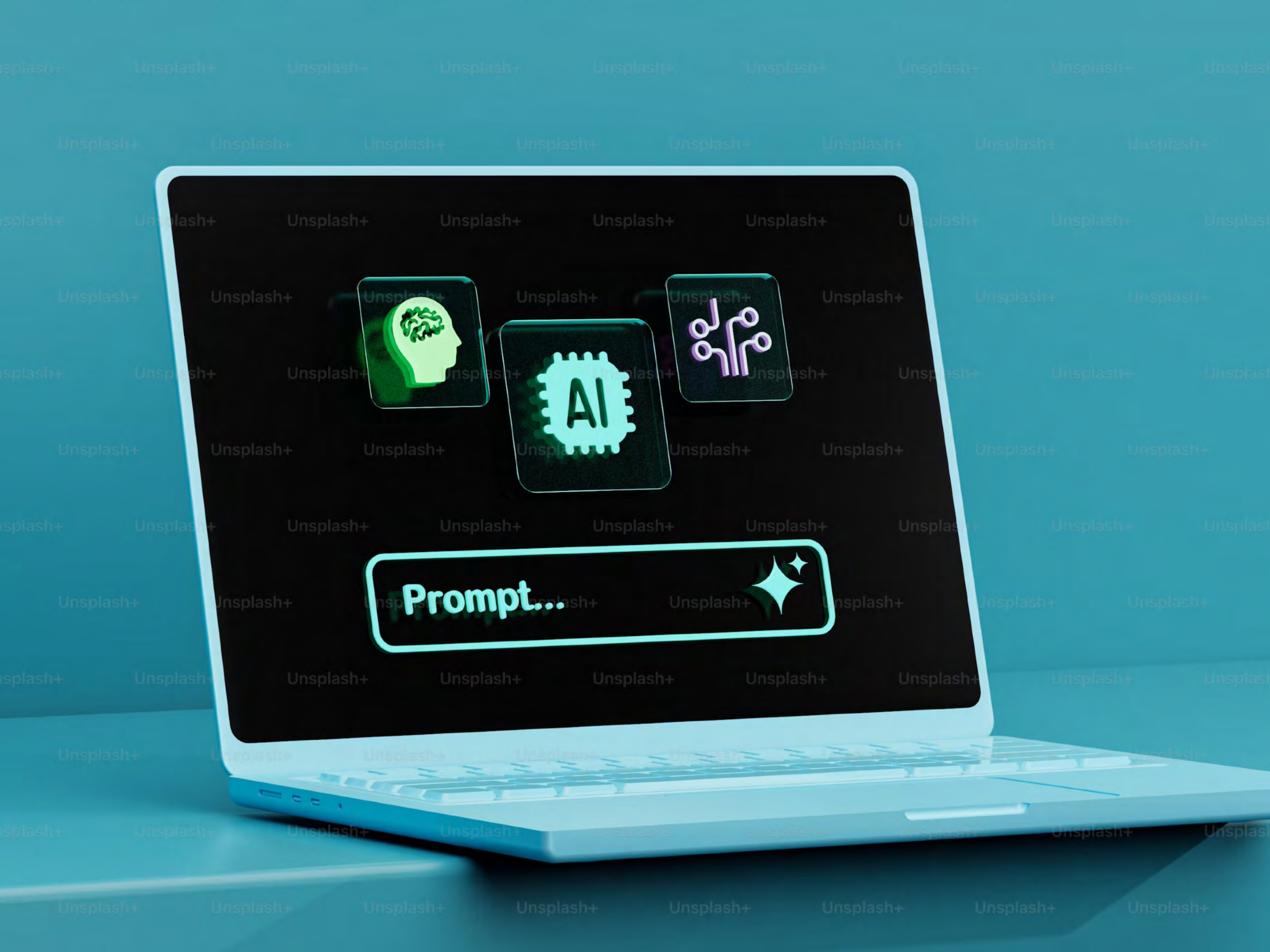
Bilingual is
90%
(vs 53% in 2008)
of Monolingual

Ad-Hoc TEL Monolingual German Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision



Ad-Hoc TEL Bilingual German Task Top 5 Participants – Standard Recall Levels vs Mean Interpolated Precision





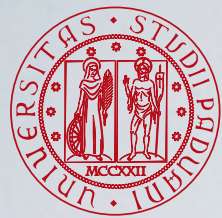
R.I.P. Information Retrieval?



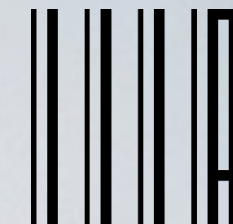


Retrieval Augmented Generation

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-t., Rocktaschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Proc. 34th Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020.



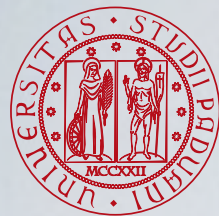
Do We Always Need Generation?



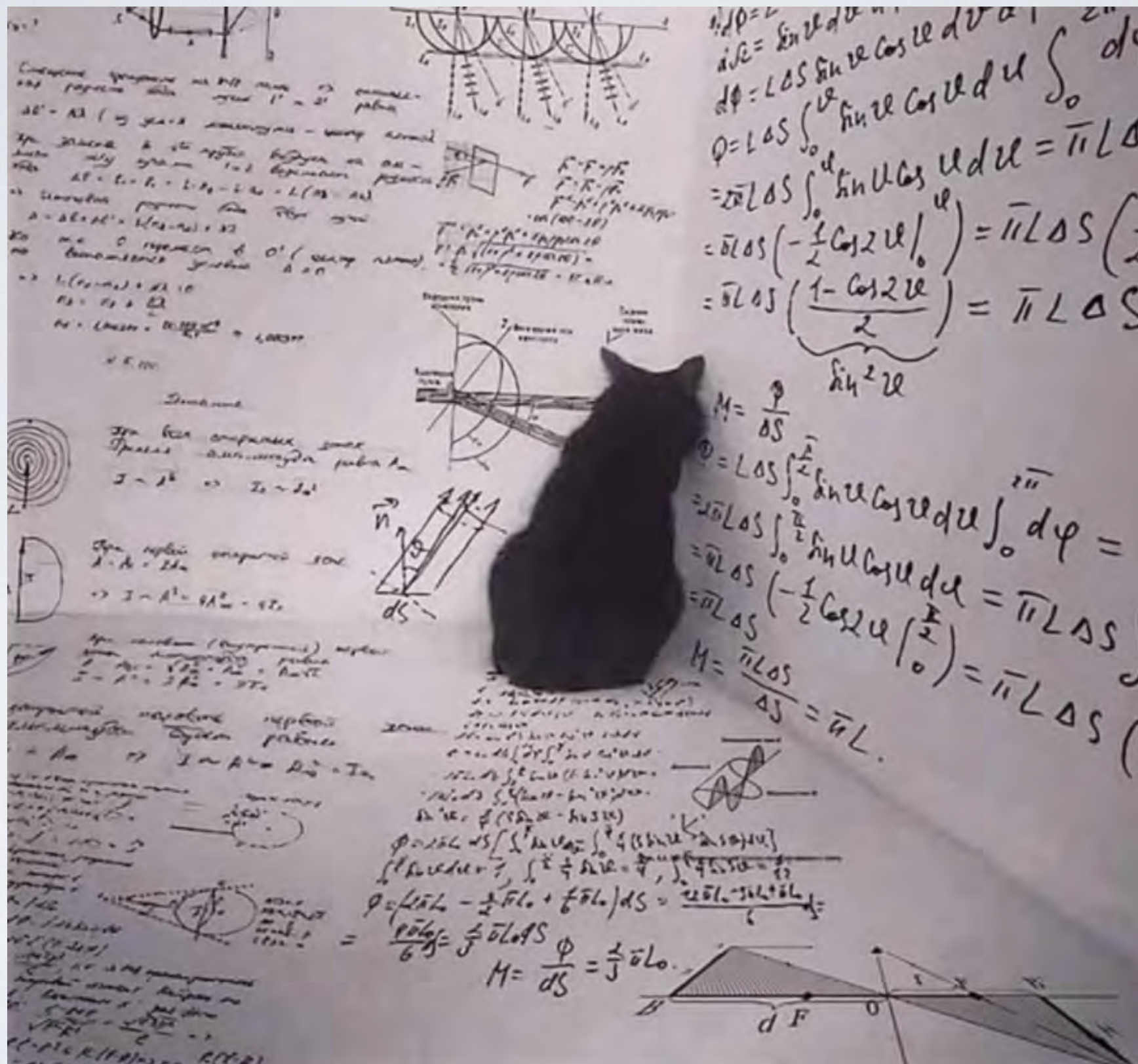
Generation



Retrieval



How to Adapt Cranfield for Evaluating RAG?



THANK YOU