The 2003 CL-SDR Track: Overview

Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica

I-38050 Povo (Trento), Italy

federico@itc.it

Introduction

Motivations:

- **Relevance of digital audio/video archives in the digital library landscape.**
- **Progress of automatic speech recognition for the purpose of indexing.**
- **Success in TREC with Spoken Document Retrieval (SDR) of American broadcast news**

Objective:

- **To investigate Cross-Language SDR (CL-SDR) with European languages.**

Constraints:

- **No fundings: for data collection and track organization.**

CLEF CL-SDR Benchmark

During 2002 ITC-irst prepared the following benchmark, mainly based on existing resources:

- **Document collection (NIST)**
  - **– automatic transcripts (75% accuracy) of 389h of American English broadcast news**
  - **– corresponding to 21,754 manually segmented stories**
- **Topics and relevant documents (NIST)**
  - **– 49+50 short topics in English used in TREC'99 and TREC'00, respectively**
  - **– 1818 + 2216 relevant documents**
- **Translations of topics (ITC-irst)**
  - **– available for Italian, Dutch, German, French and Spanish**
- **Additional documents (distributed by LDC):**
  - **– 314,697 texts North American News (Sept. '97 - Apr. '98)**

CLEF CL-SDR Track

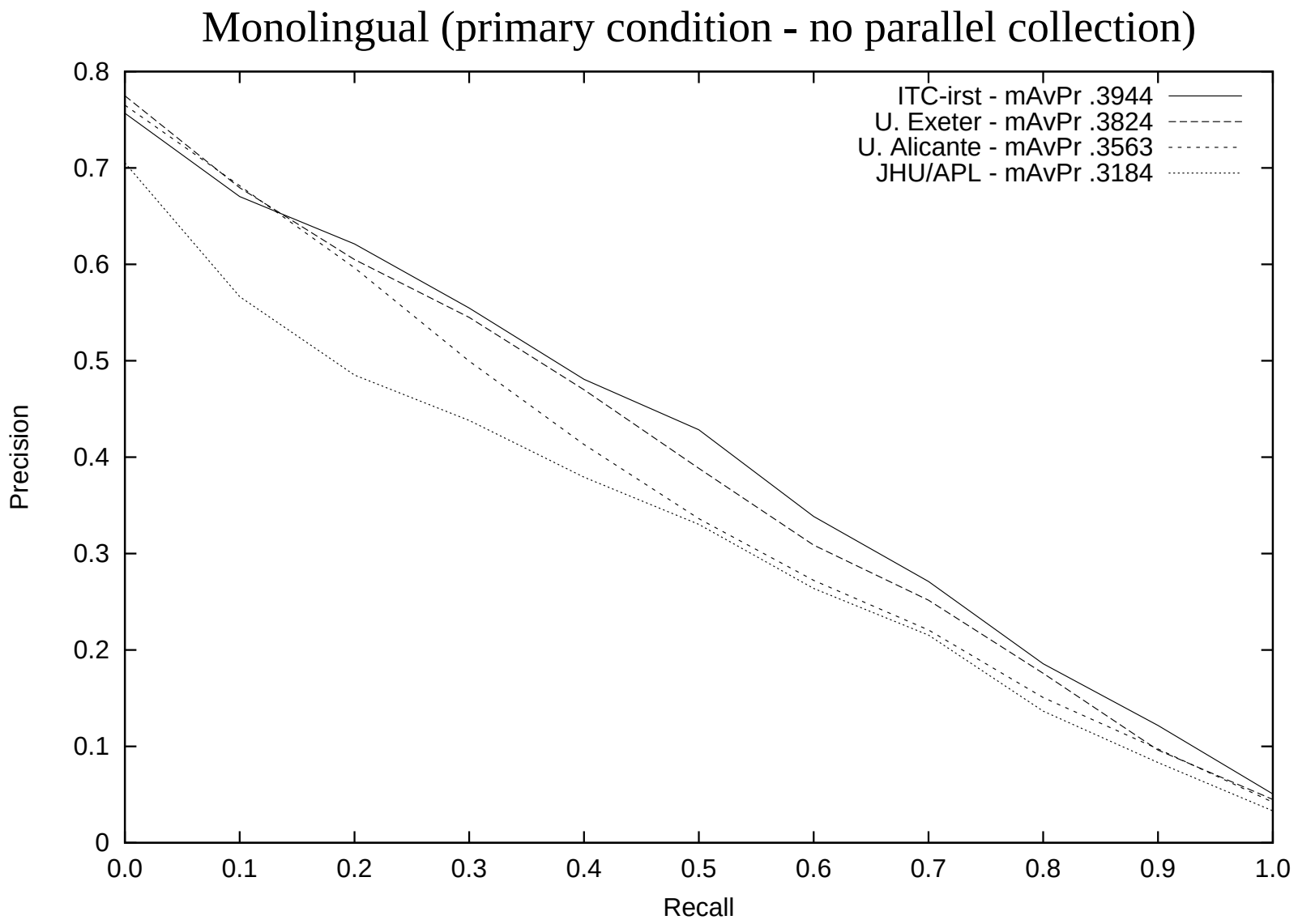
- Development data: TREC ’99 collection
- Evaluation data: TREC ’00 collection
- Primary condition (mandatory)
  - Bilingual French-English or German-English
  - Monolingual English with no parallel collection (contrastive condition)
- Secondary condition (optional)
  - Bilingual Italian-English, Spanish-English, Dutch-English

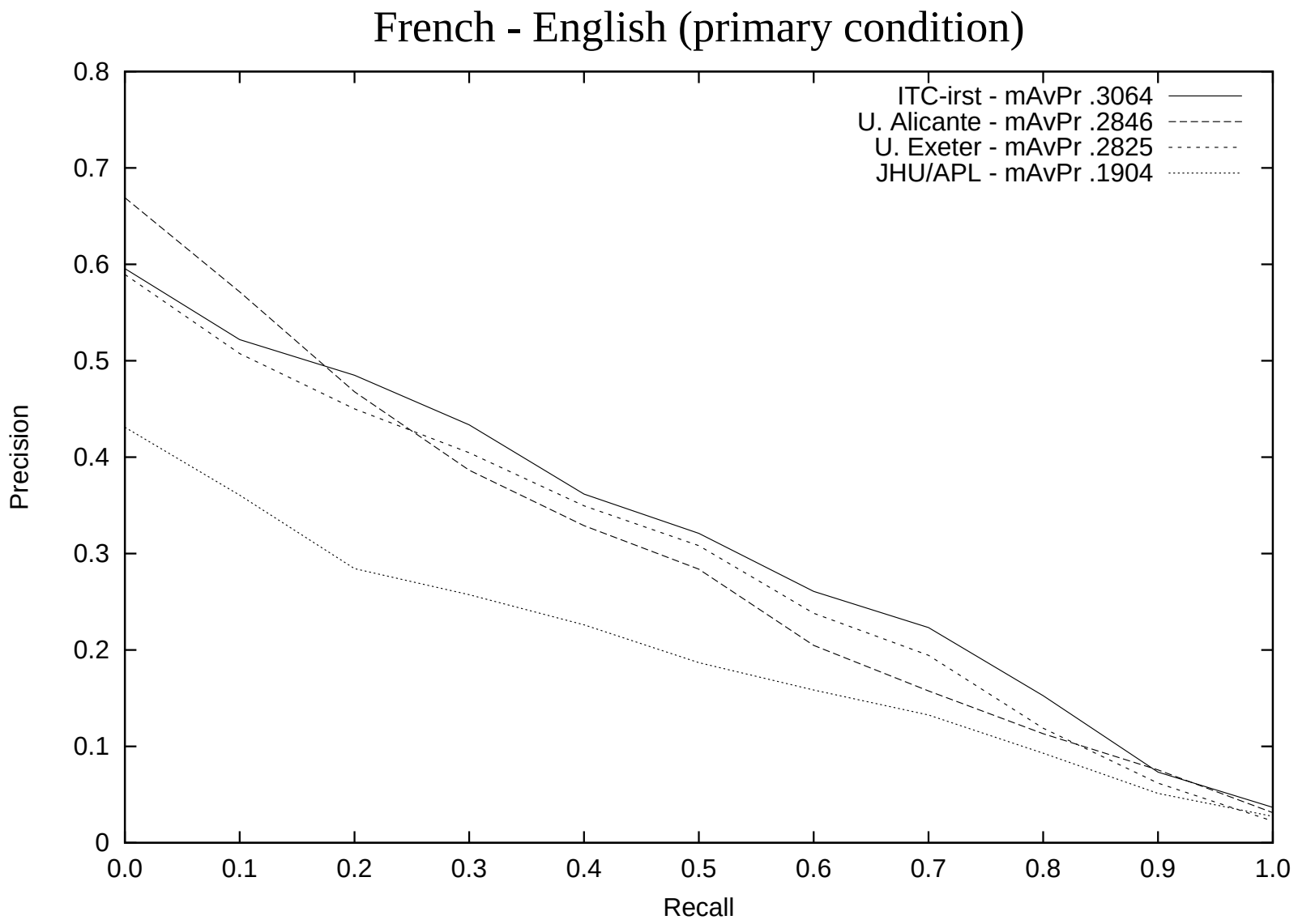
More details in the track WEB page: `http://munst.itc.it/clef-sdr.html`

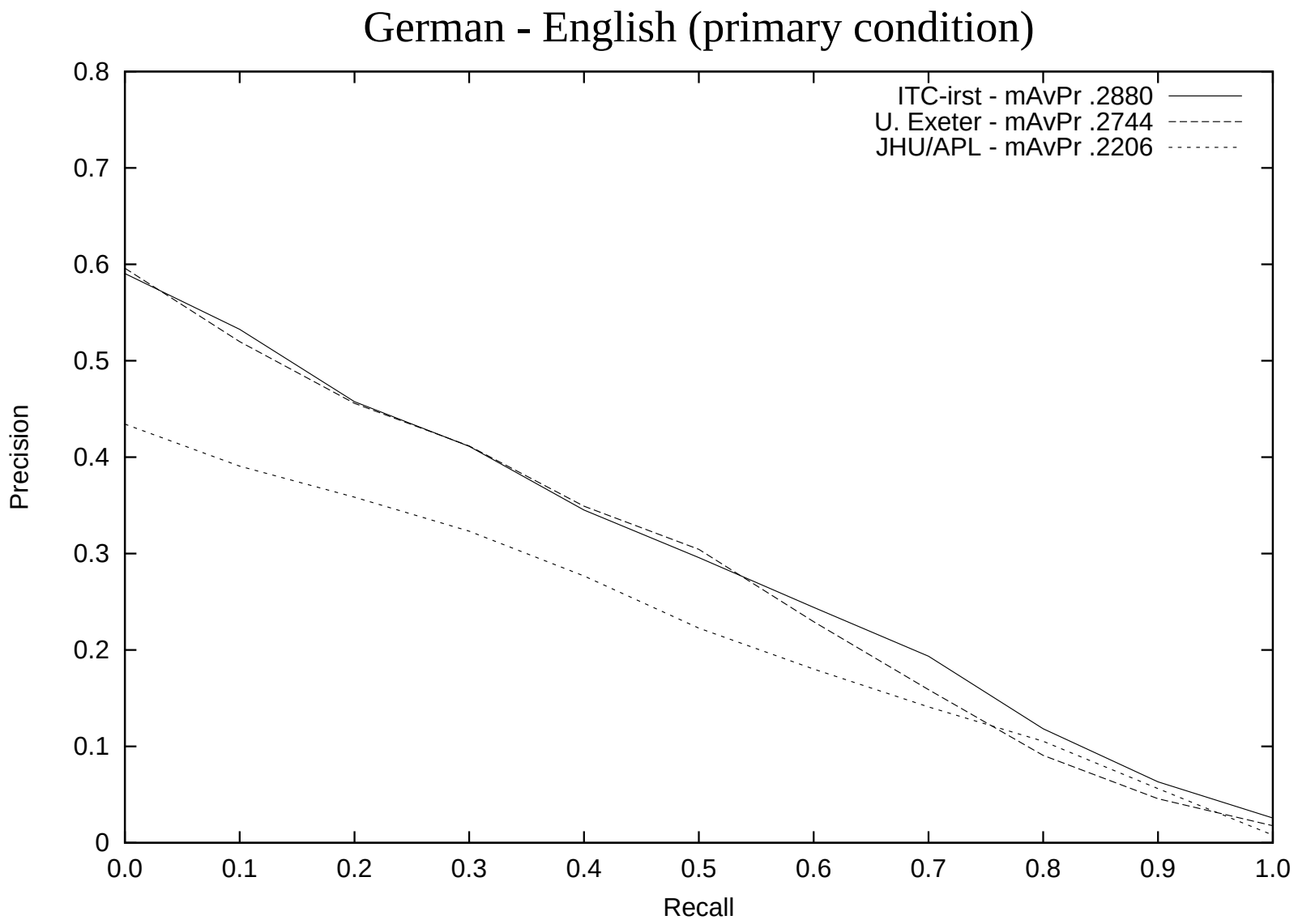
Participants and submitted runs

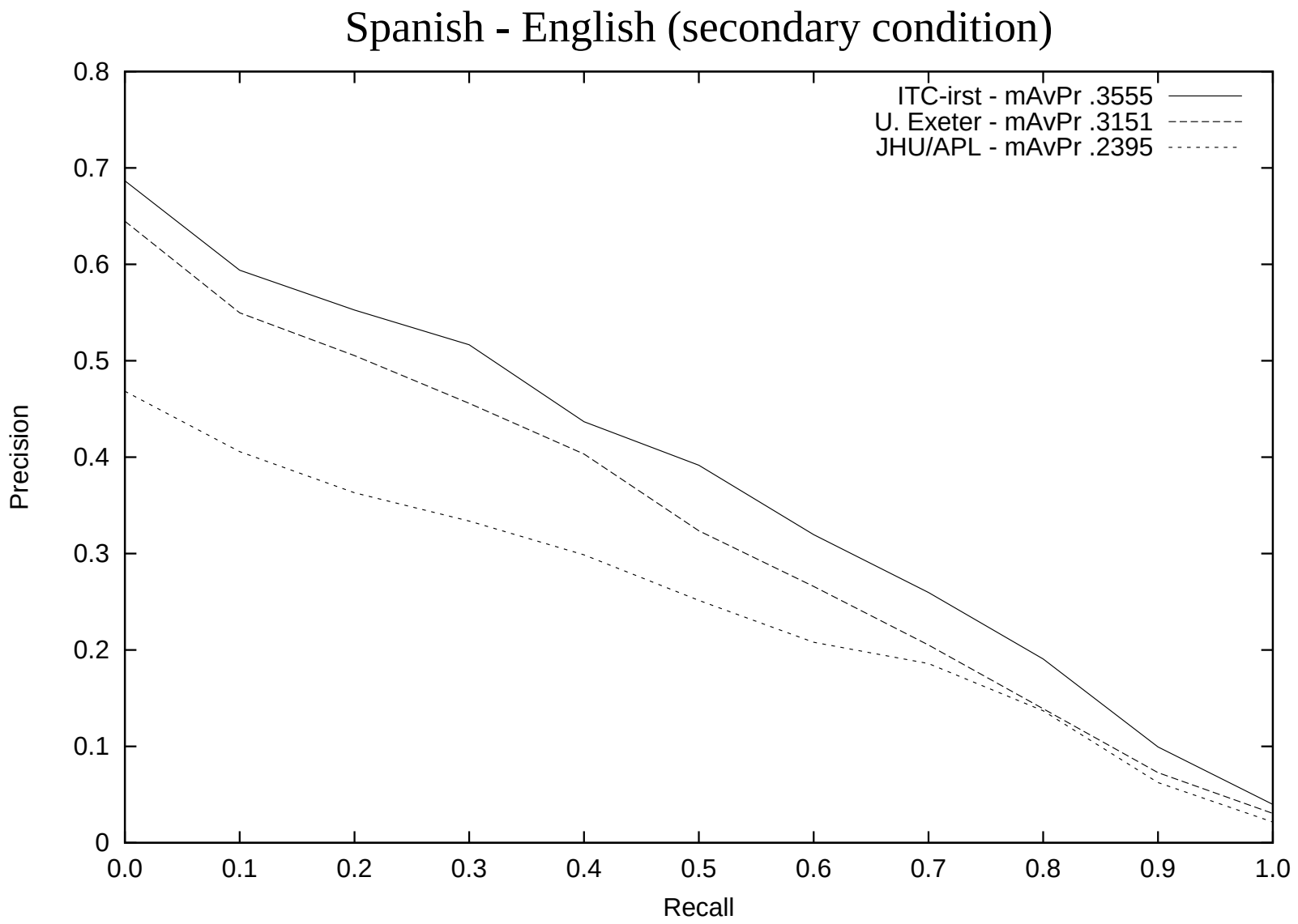
- University of Alicante (Spain): 2 x (EN+FR)
- Johns Hopkins University (USA): 1 x (EN+FR+DE+IT+ES+NL)
- University of Exeter (U.K.): 2 x (EN+FR+DE+IT+ES)
- ITC-irst (Italy): 1 x EN + 2 x (FR+DE+IT+ES)

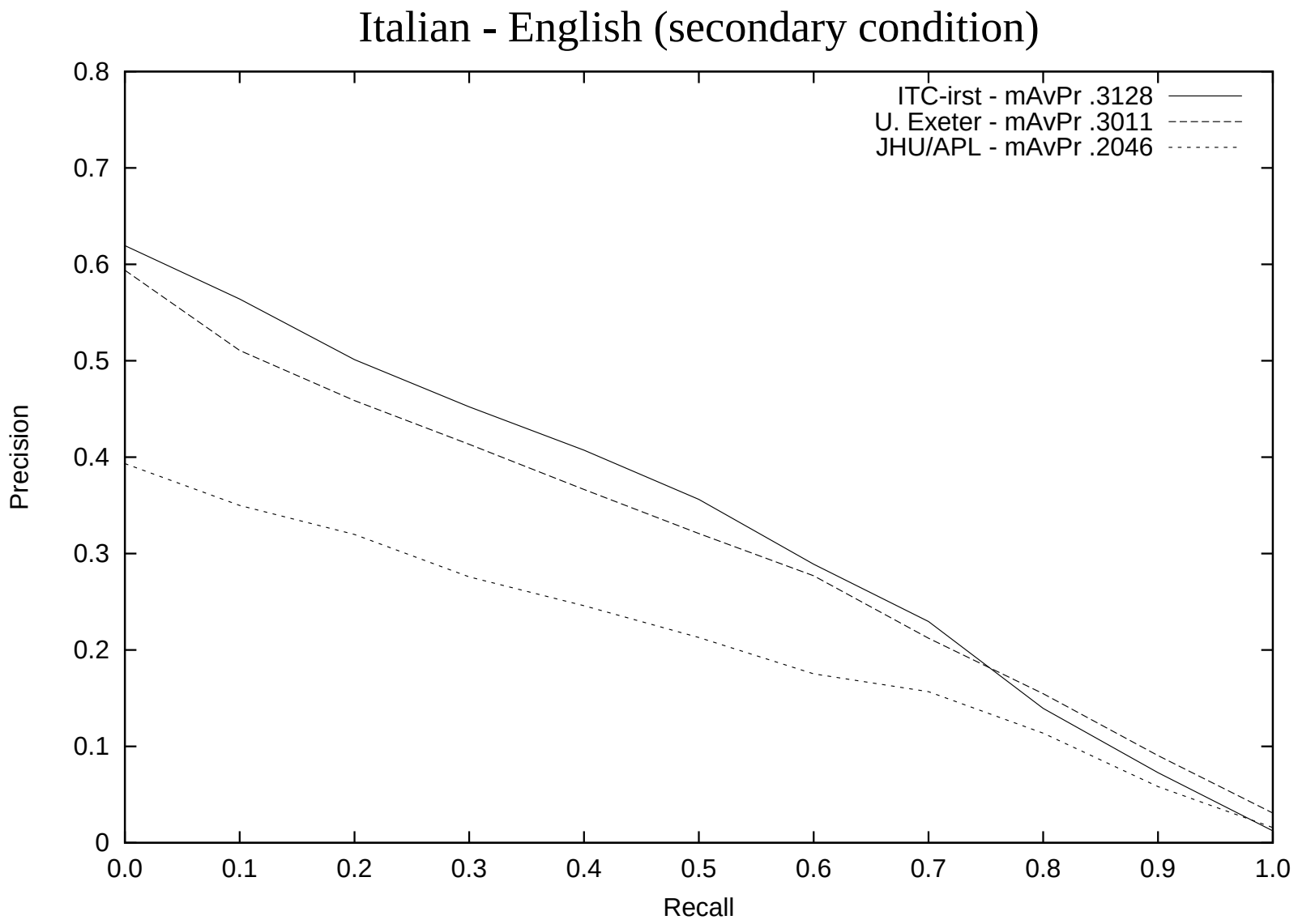
Remark: just one submitted run for Dutch-English.

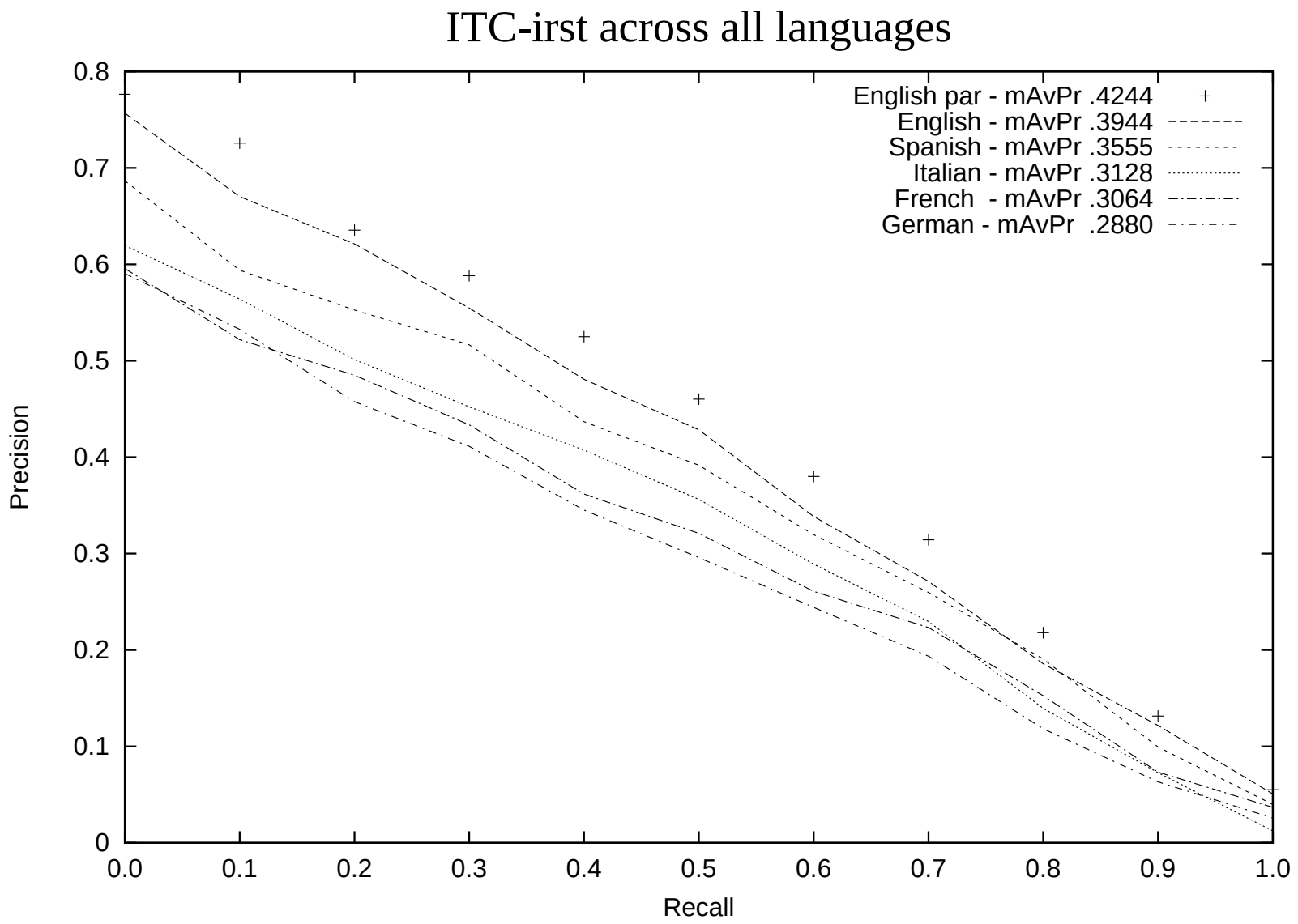

Monolingual (primary condition - no parallel collection)
ITC-irst - mAvPr .3944
U. Exeter - mAvPr .3824
U. Alicante - mAvPr .3563
JHU/APL - mAvPr .3184
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

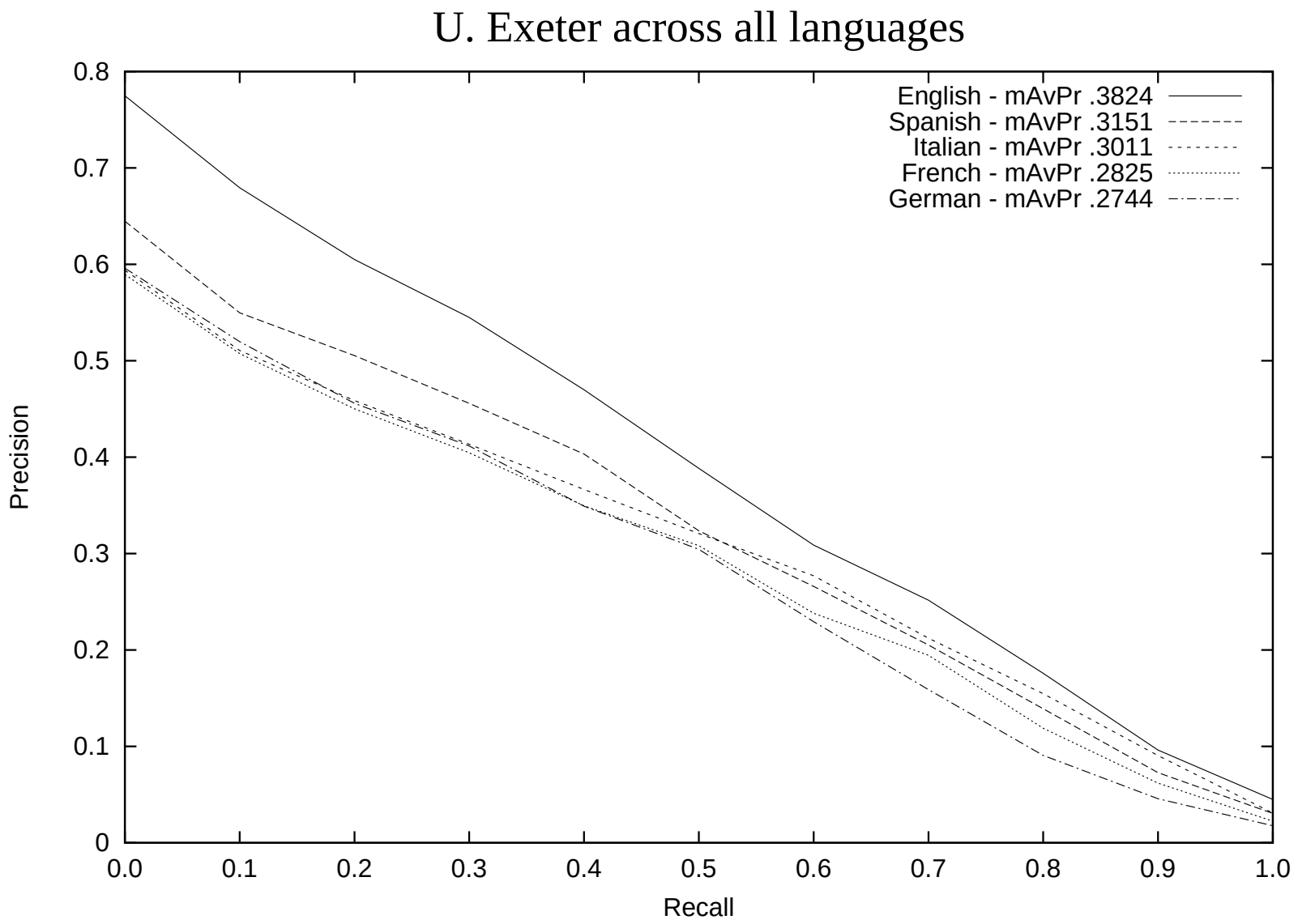

French - English (primary condition)
ITC-irst - mAvPr .3064
U. Alicante - mAvPr .2846
U. Exeter - mAvPr .2825
JHU/APL - mAvPr .1904
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

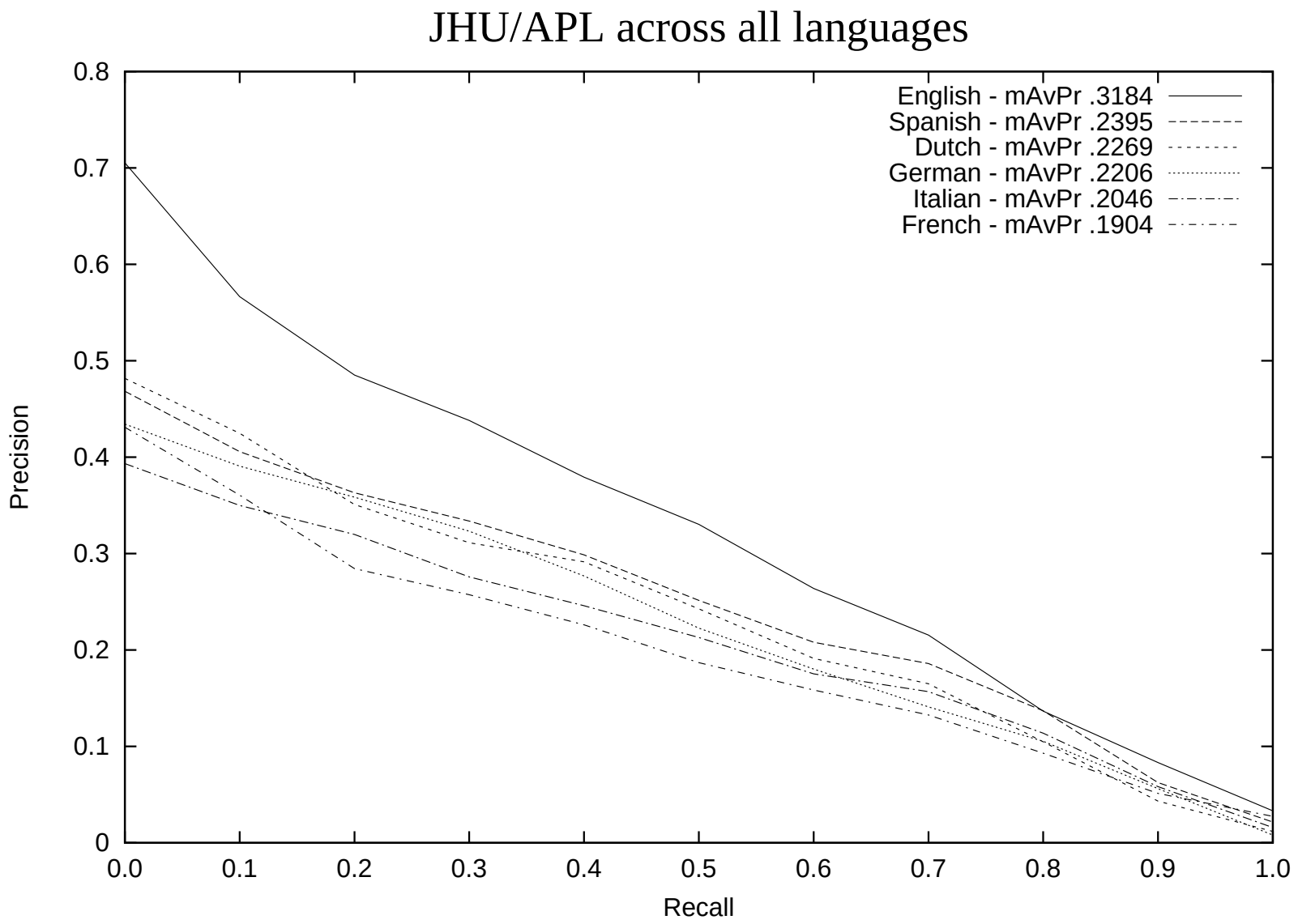

German - English (primary condition)
ITC-irst - mAvPr .2880
U. Exeter - mAvPr .2744
JHU/APL - mAvPr .2206
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


Spanish - English (secondary condition)
ITC-irst - mAvPr .3555
U. Exeter - mAvPr .3151
JHU/APL - mAvPr .2395
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

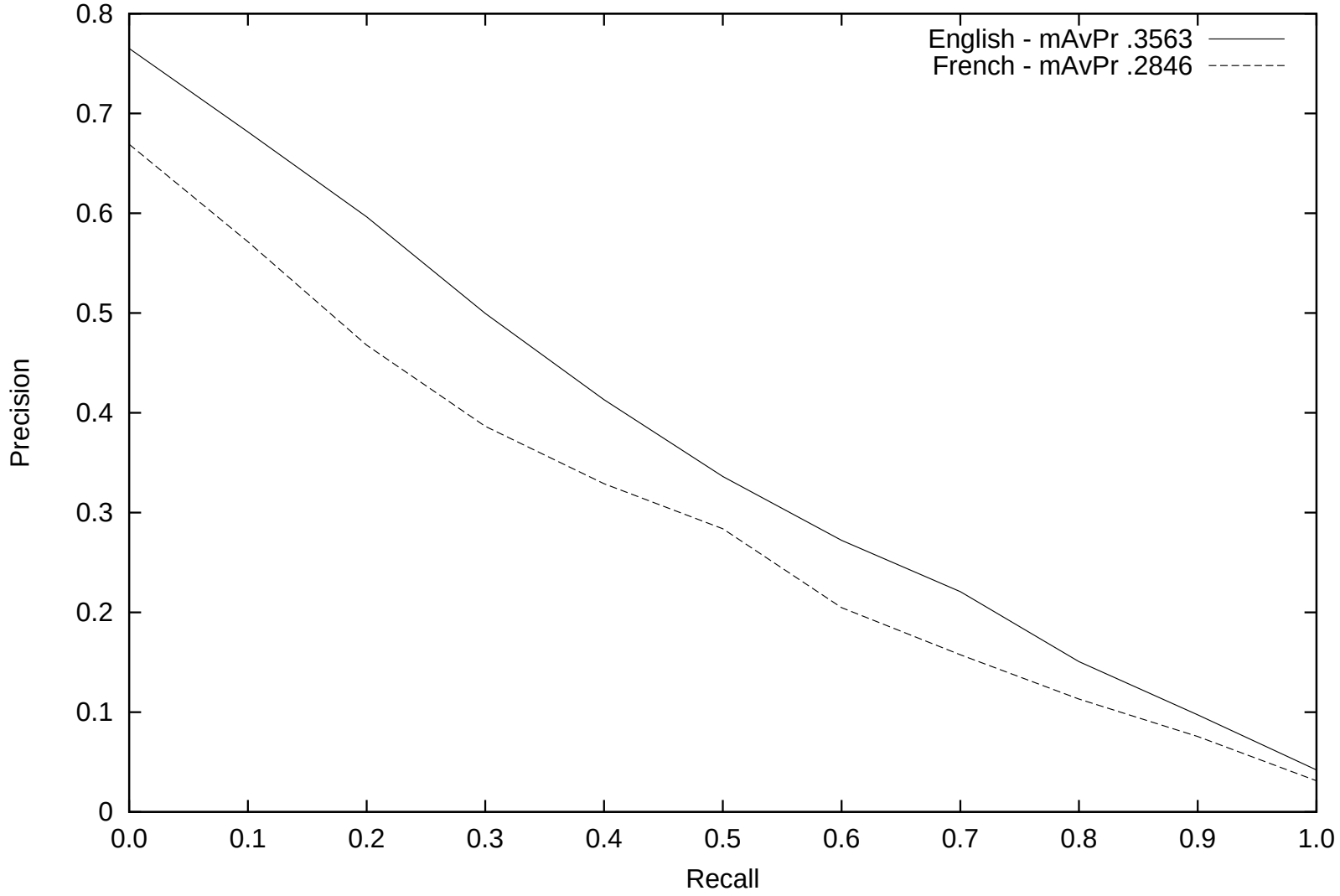

Italian - English (secondary condition)
ITC-irst - mAvPr .3128
U. Exeter - mAvPr .3011
JHU/APL - mAvPr .2046
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


ITC-irst across all languages
English par - mAvPr .4244
English - mAvPr .3944
Spanish - mAvPr .3555
Italian - mAvPr .3128
French - mAvPr .3064
German - mAvPr .2880
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


U. Exeter across all languages
English - mAvPr .3824
Spanish - mAvPr .3151
Italian - mAvPr .3011
French - mAvPr .2825
German - mAvPr .2744
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


JHU/APL across all languages
English - mAvPr .3184
Spanish - mAvPr .2395
Dutch - mAvPr .2269
German - mAvPr .2206
Italian - mAvPr .2046
French - mAvPr .1904
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

U. Alicante across all languages

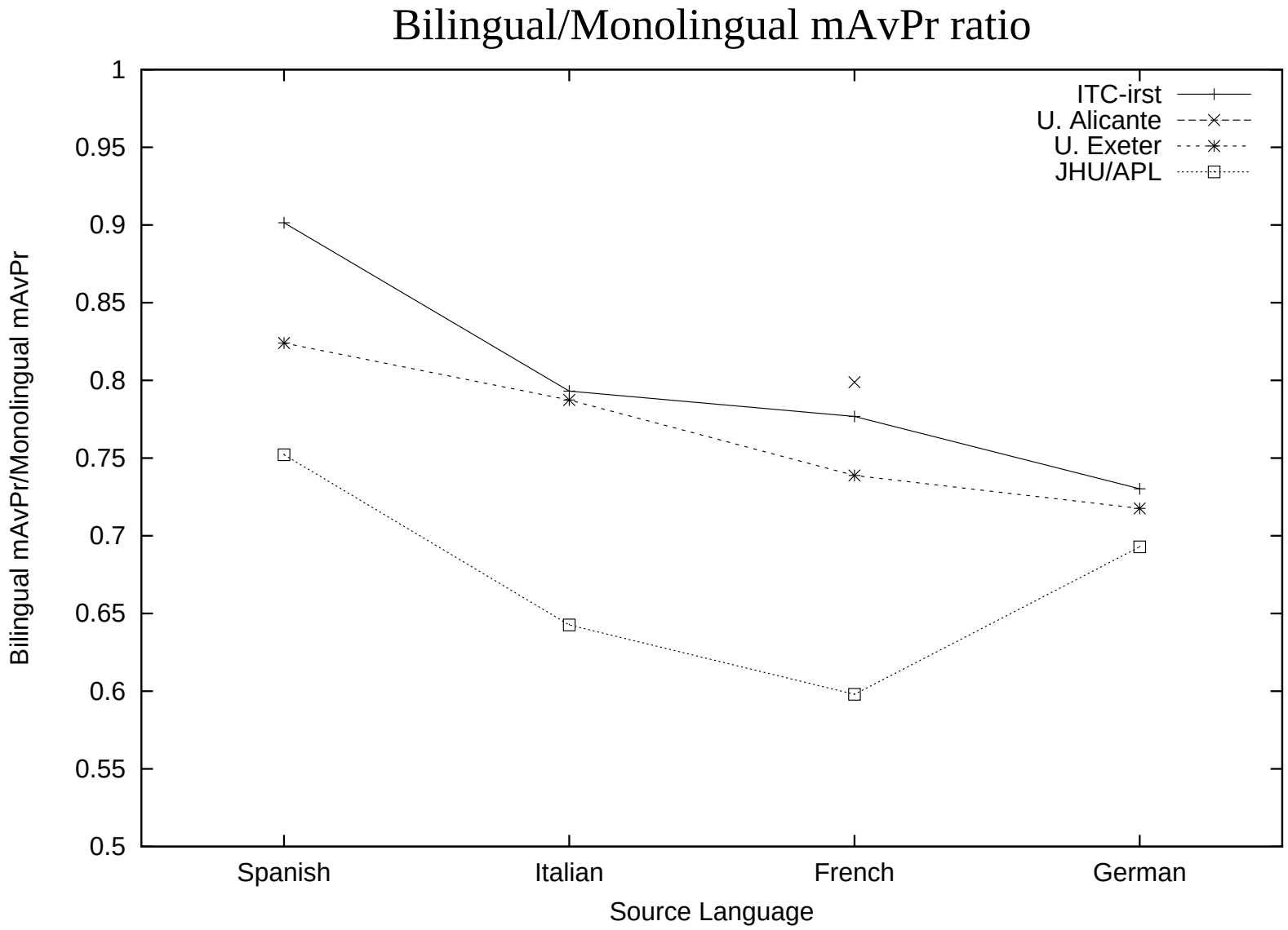


Bilingual/Monolingual mAvPr Ratio

The ratio between bilingual `mAvPr` and monolingual `mAvPr` can be viewed as:

- **a normalized bilingual-retrieval performance**
  - **– assumption: we cannot do better than the monolingual case**
- **a measure of effectiveness of the 'cross-lingual' components**
  - **– e.g. query translation component**

In the following, slide we compare this measure across all systems and languages.


Bilingual/Monolingual mAvPr ratio
Bilingual mAvPr/Monolingual mAvPr
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Spanish
Italian
French
German
Source Language
ITC-irst
U. Alicante
U. Exeter
JHU/APL

Conclusions

- The CL-SDR track was basically a bilingual document retrieval track featuring:
  - short queries, which seems to make translation more difficult
  - a small collection of noisy texts, which affects ordinary relevance feedback techniques
- Loss in performance w.r.t. to monolingual IR was similar to that for clean texts
  - $\approx$ 15% in the best cases (English-Spanish)
- Contrastive condition provided some insights about systems
  - why not applying this to other tracks, too?
- There is still much room for improvement
  - current performance is not satisfactory
- The SDR TREC data still presents many interesting issues to cope with
  - hence, we believe there is a future for this track

Future Work

Ideas about the next CL-SDR tracks:

- **Unknown story boundary condition:**
  - **move to the real task: retrieval of relevant segments of transcripts**
- **Multiple transcription hypotheses:**
  - **try to exploit alternative transcriptions provided by different speech recognizers**
- **Short queries and small collection:**
  - **improve translation and retrieval techniques for this specific condition**
- **Informative/reliable evaluation:**
  - **define contrastive conditions to evaluate single components, e.g. fix translations**
  - **check components on different development-test configurations to reduce bias**
- **Cooperative development:**
  - **one site could focus on some component and let the others test it in their systems**