CLEF 2001
Overview of Results

Martin Braschler
Eurospider Information Technology AG
8006 Zürich, Switzerland
braschler@eurospider.com

Outline

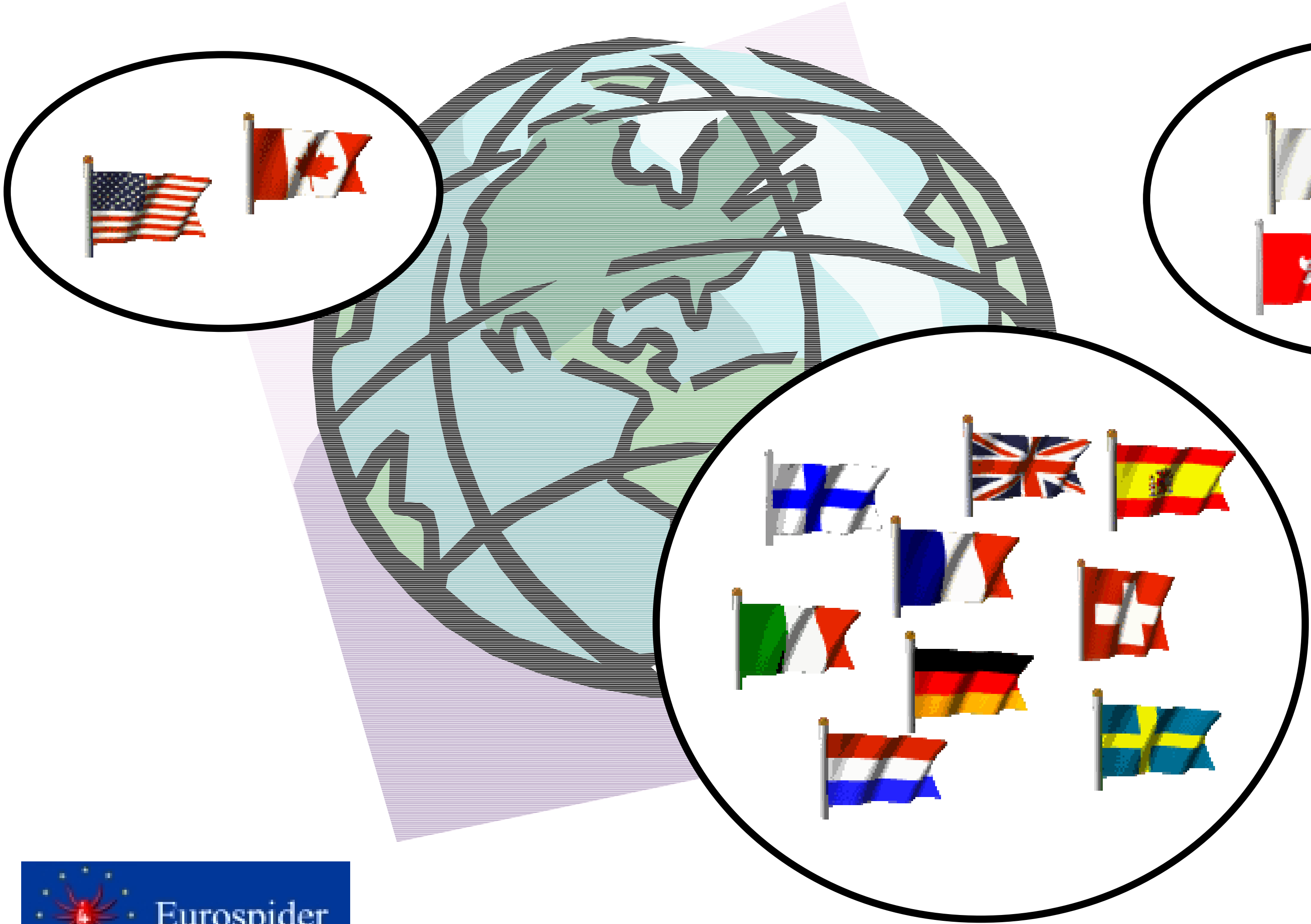
- Participants
- Experiment Details
- Trends & Effects
- Results

Participants

- CMU
- Eidetica
- Eurospider *
- Greenwich U
- HKUST
- Hummingbird
- IAI *
- IRIT *
- ITC-irst *
- JHU-APL *
- Kasetsart U
- KCSL Inc.
- Medialab
- Nara Inst. of Tech.
- National Taiwan U
- OCE Tech. BV
- SICS/Conexor
- SINAI/U Jaen
- Thomson Legal *
- TNO TPD *
- U Alicante
- U Amsterdam
- U Exeter
- U Glasgow *
- U Maryland * (interactive only)
- U Montreal/RALI *
- U Neuchâtel
- U Salamanca *
- U Sheffield * (interactive only)
- U Tampere *
- U Twente (*)
- UC Berkeley (2 groups) *
- UNED (interactive only)

31+3 participants, 15 different countries.
(* = also participant in 2000)

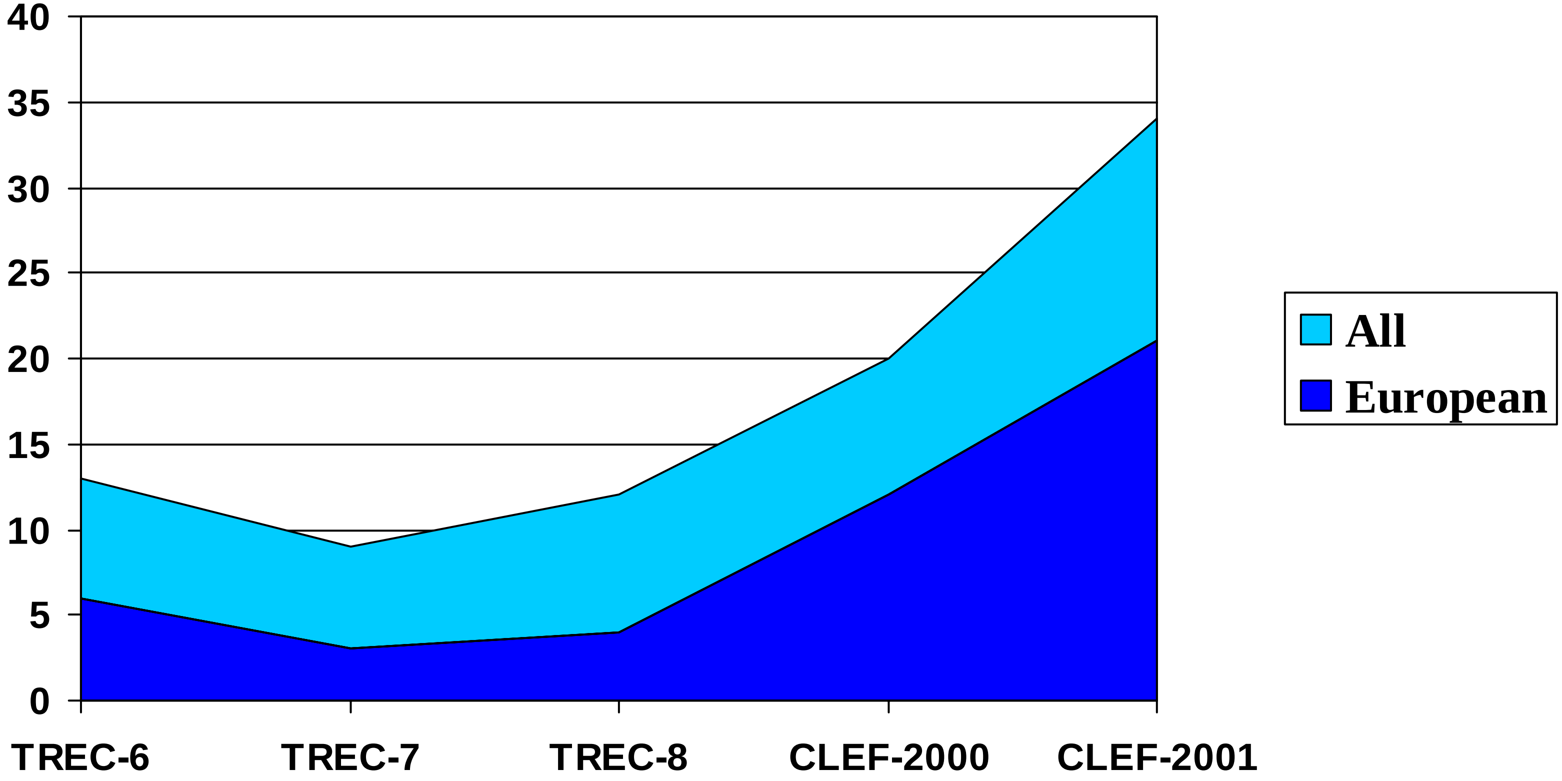
CLEF Goes Global



Flags: www.fg-a.com

CLEF Growth

(Number of Participants)



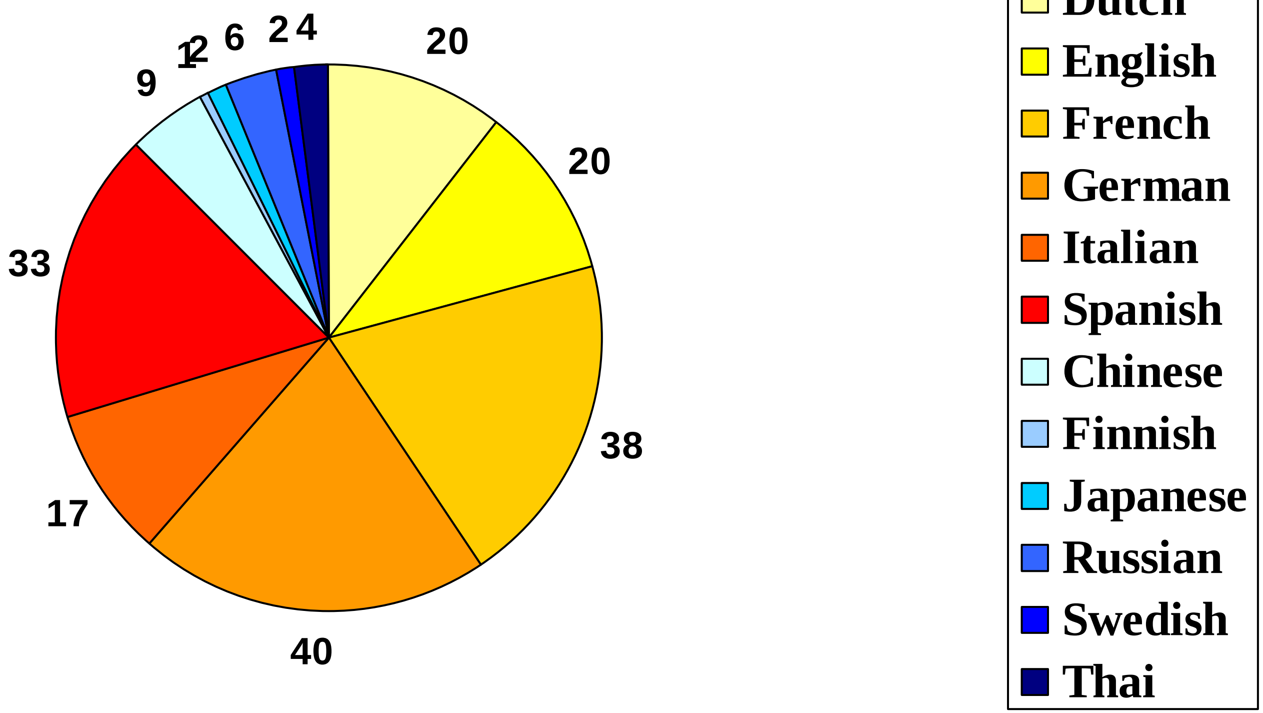
The CLEF Multilingual Collection

	# part.	# documents	Size in MB	# assess.	# topics	# ass. per topic
CLEF 2001	31	749,883	1982	80,624	50	1612
CLEF 2000	20	368,763	1158	43,566	40	1089
TREC8 CLIR	12	698,773	1620	23,156	28	827
TREC8 AdHoc	41	528,155	1904	86,830	50	1736
TREC7 AdHoc	42+4	528,155	1904	~80,000	50	~1600

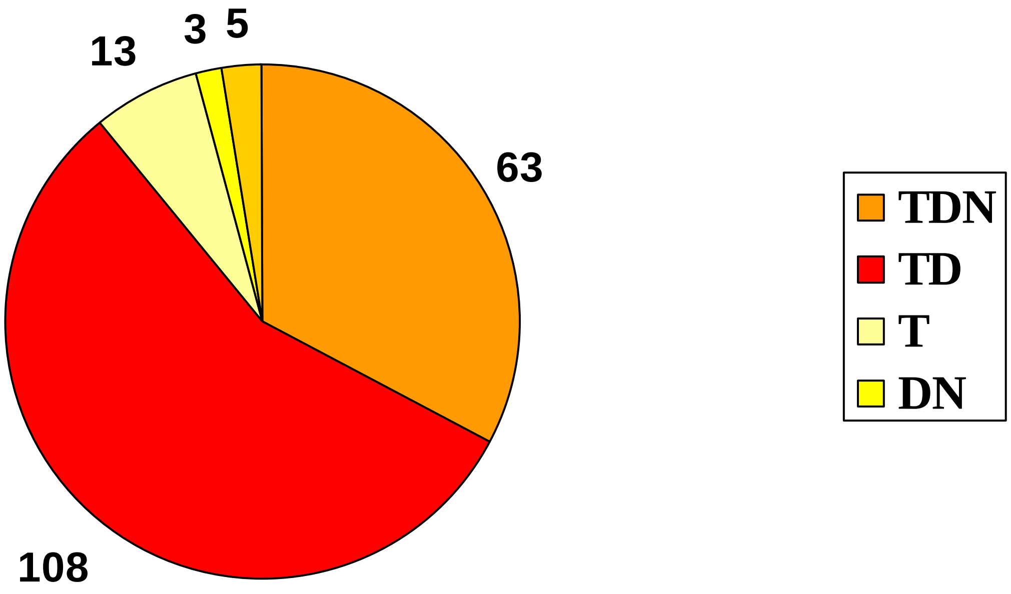
Details of Experiments

Track	# Participants	# Runs/Experiments
Multilingual	8	26
Bilingual to EN	19	61
Bilingual to NL	3	3
Monolingual DE	12	25
Monolingual ES	10	22
Monolingual FR	9	18
Monolingual IT	8	14
Monolingual NL	9	19
Domain-specific	1	4
Interactive	3	6

Runs per Topic Language



Topic Fields



Trends Observed for CLEF-2001 (1)

- Corpus-based, statistical approaches were very popular: UC Berkeley, TNO, RALI, CMU, Mercure, Nara, Eurospider, U Glasgow, Thomson Legal, ..
- Use for either translation or for disambiguation of translation alternatives
- Mining of parallel web pages (some also provided through CLEF participants)
- Use of other parallel and comparable corpora
- Often combined with MRD or MT.

Trends Observed for CLEF-2001 (2)

- A lot of work on stemming
- Interesting work on German (and Dutch, Finnish, Swedish) decompounding
- Conflicting results for decompounding:
  - no improvement: UC Berkeley (-0.5% MAP)
  - improvement: West Group (2000) (+23% MAP)
- N-grams instead of decompounding?

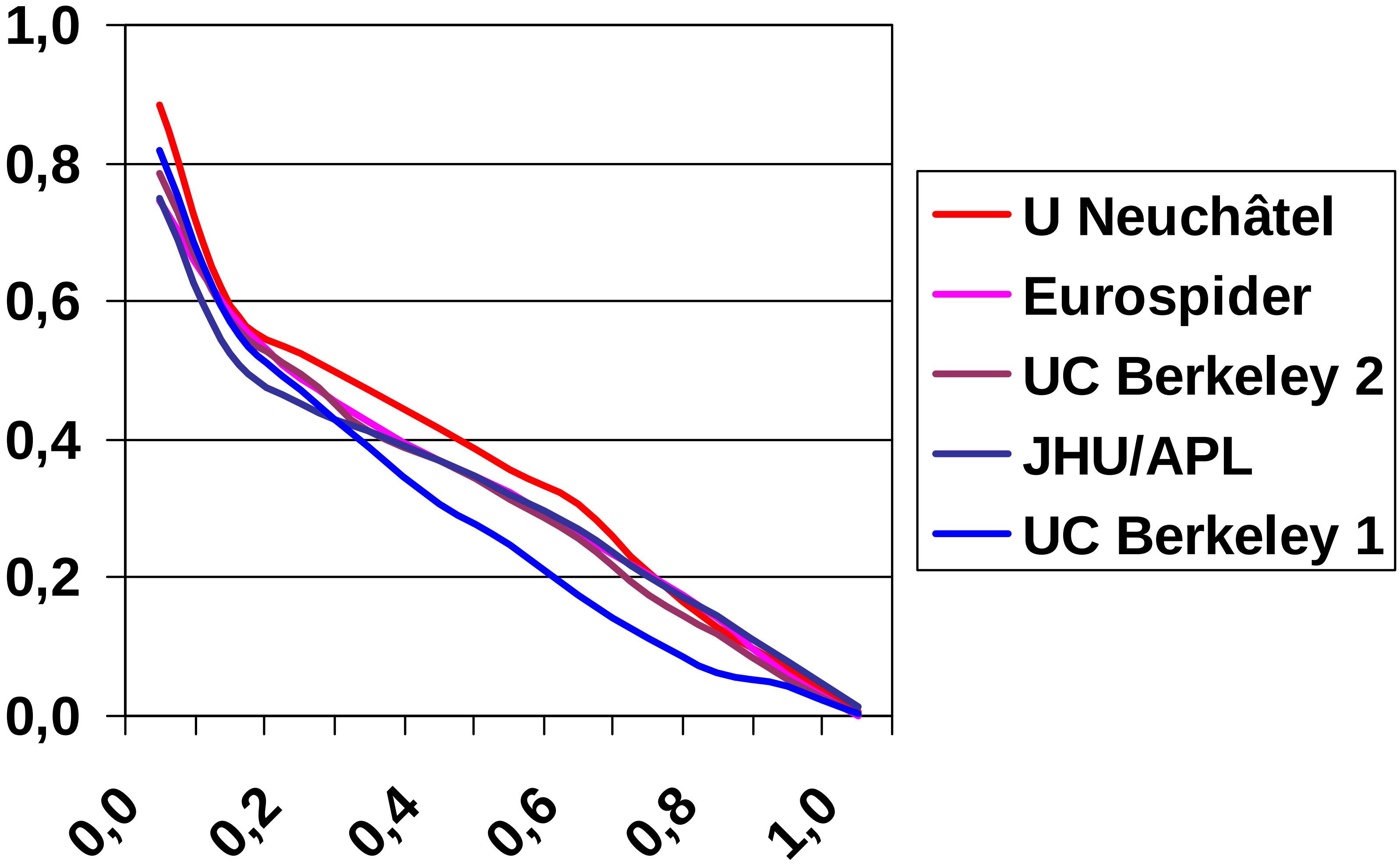
CLEF-2001 vs. CLEF-2000

- Most participants were back
- Less MT
- More Corpus-Based
- People really start to try each other’s ideas/methods:
  - corpus-based approaches (parallel web, alignments)
  - n-grams
  - combination approaches
  - etc.

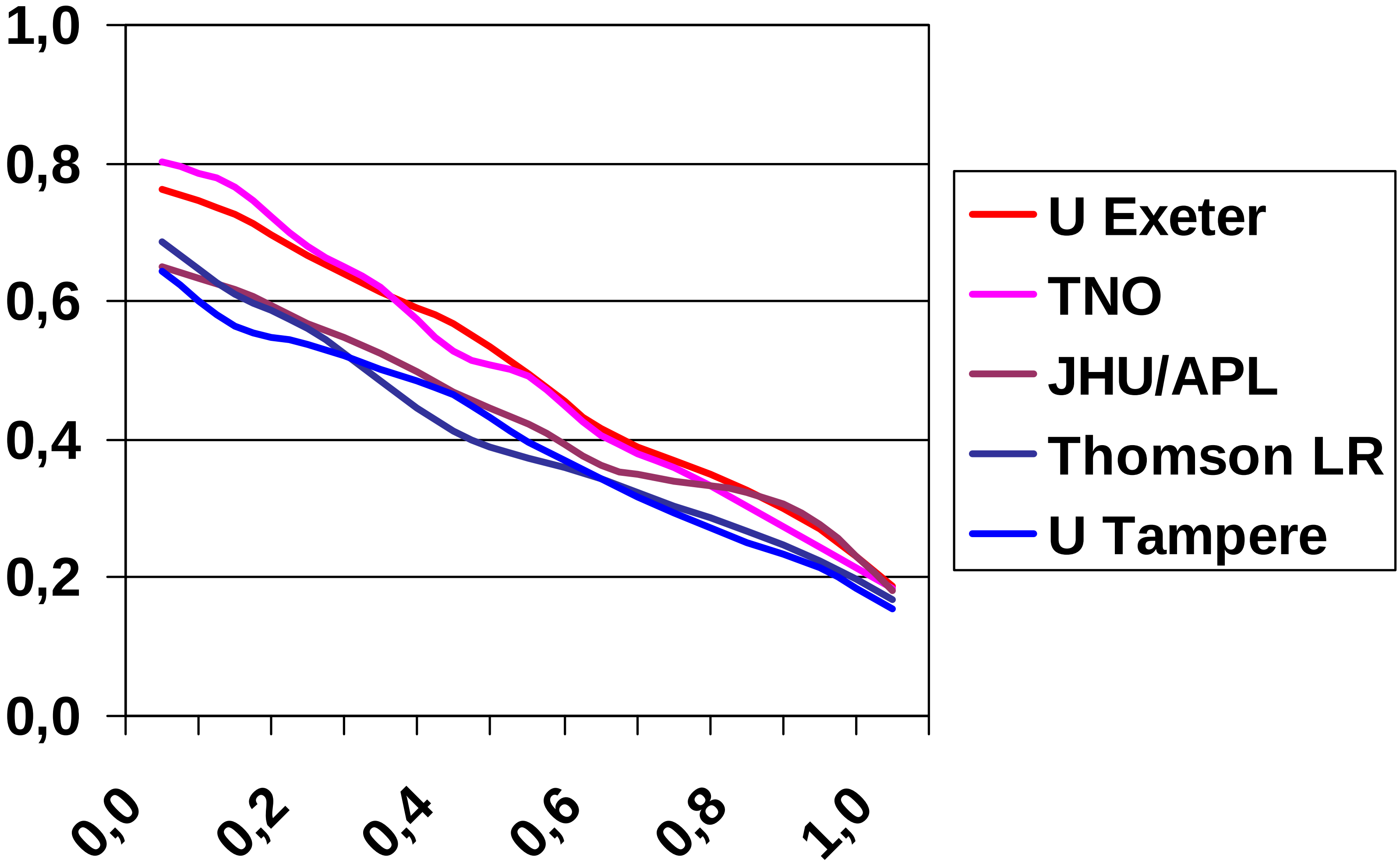
“Effect” of CLEF

- Many more European groups (21!)
- Dramatic increase of work in stemming/decompounding (for languages other than English)
- Work on mining the web for parallel texts
- Work on merging (breakthrough still missing?)
- Work on combination approaches

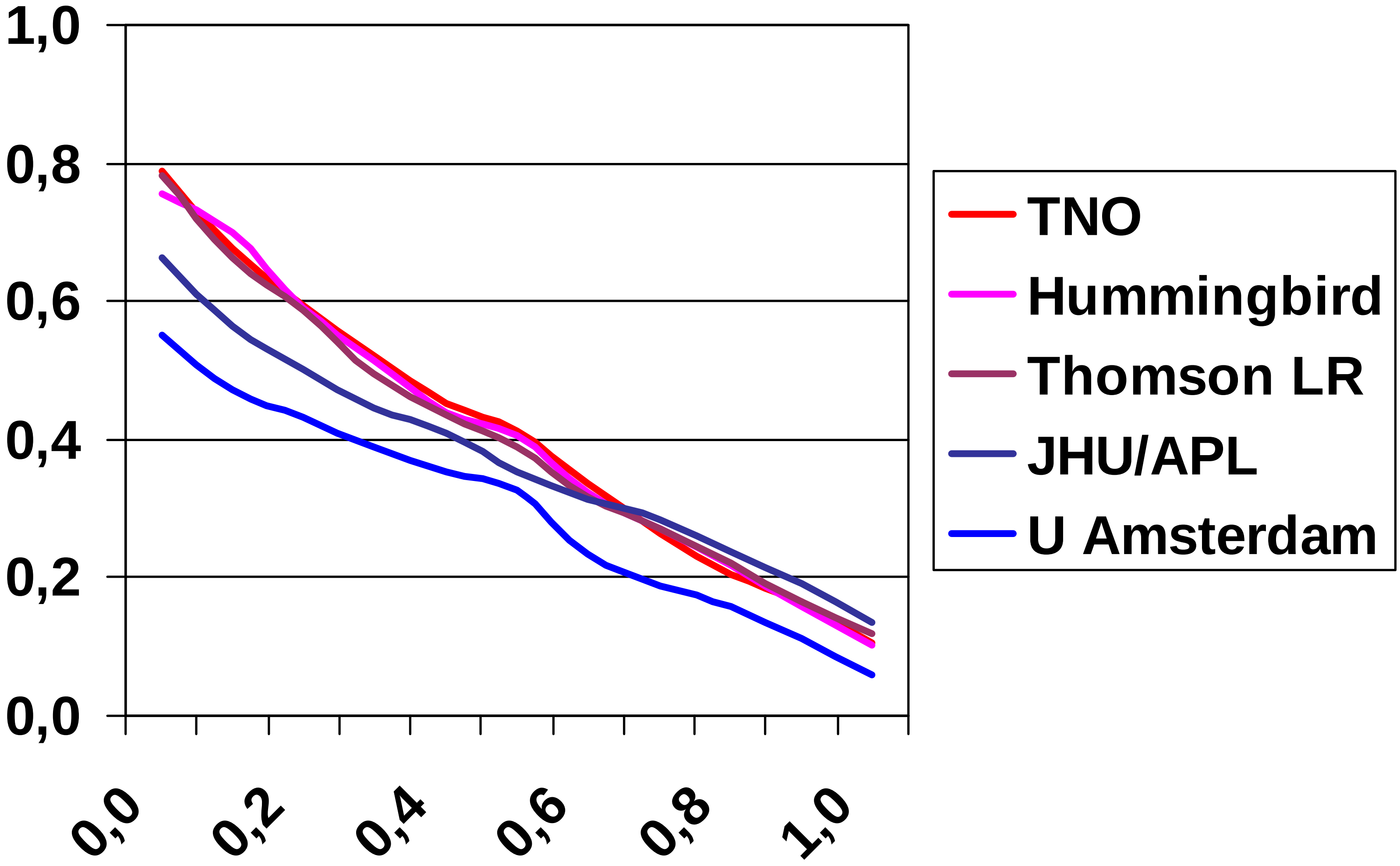
Multilingual



Bilingual

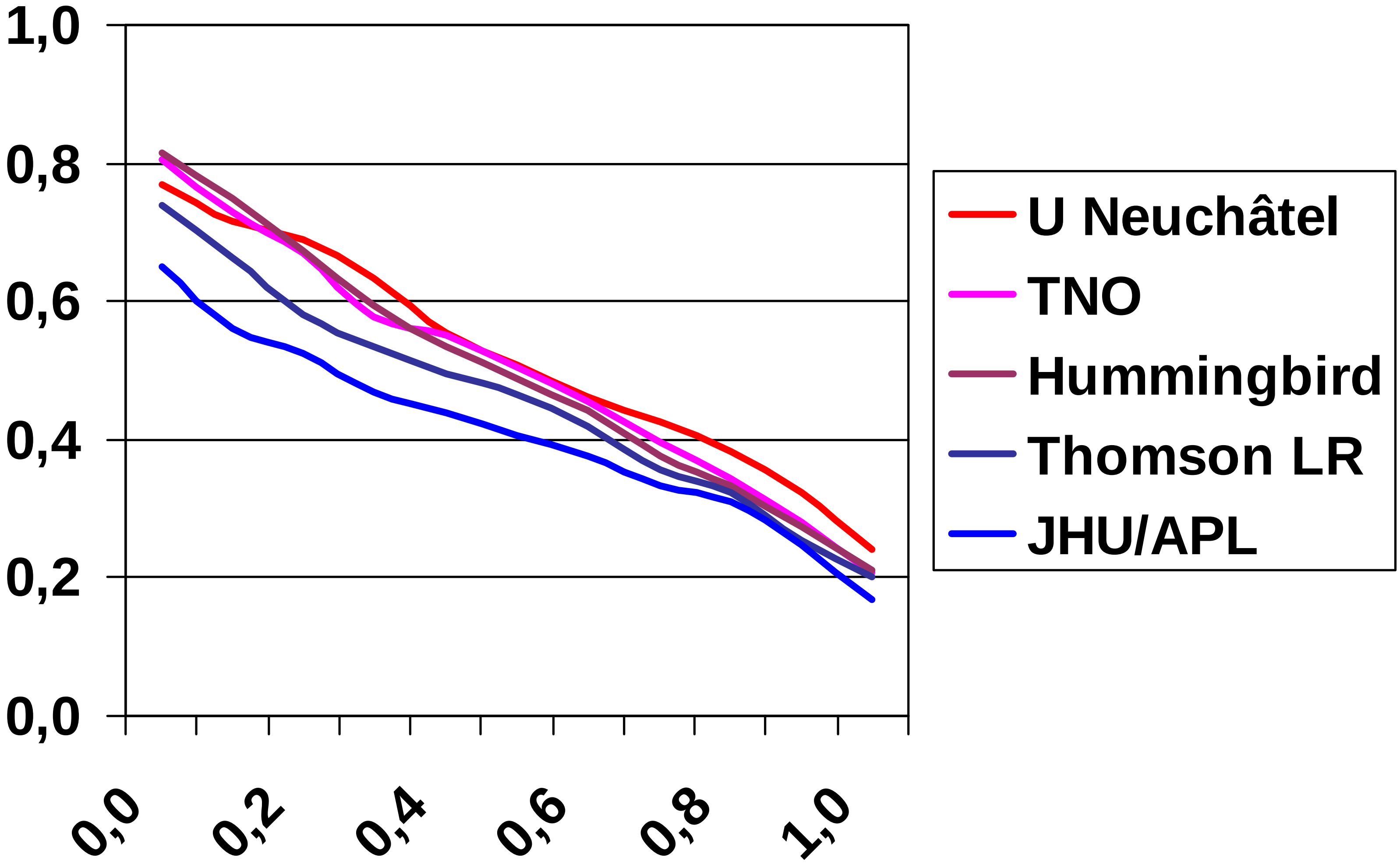

1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
U Exeter
TNO
JHU/APL
Thomson LR
U Tampere

Dutch


1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
TNO
Hummingbird
Thomson LR
JHU/APL
U Amsterdam

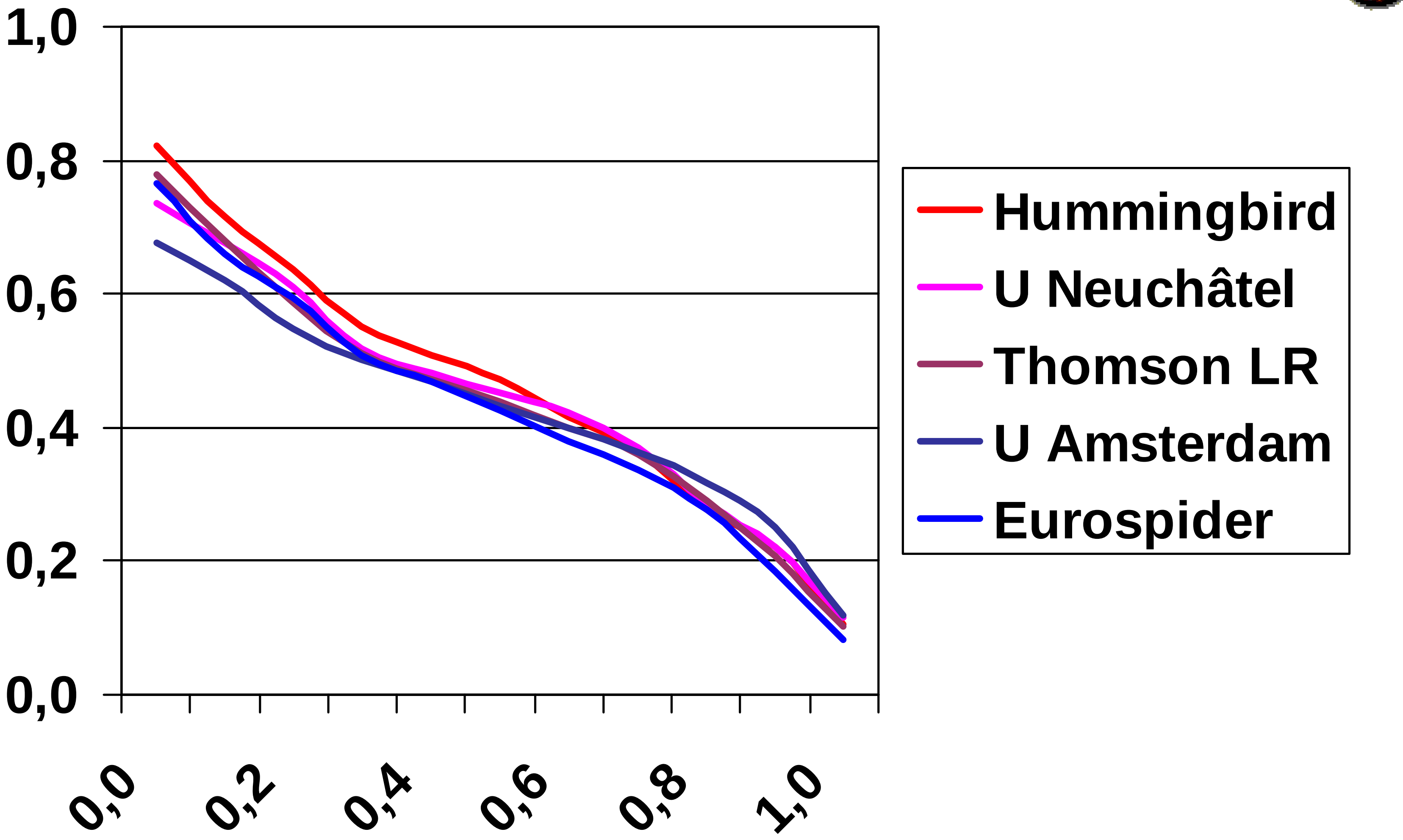
French

CLEF

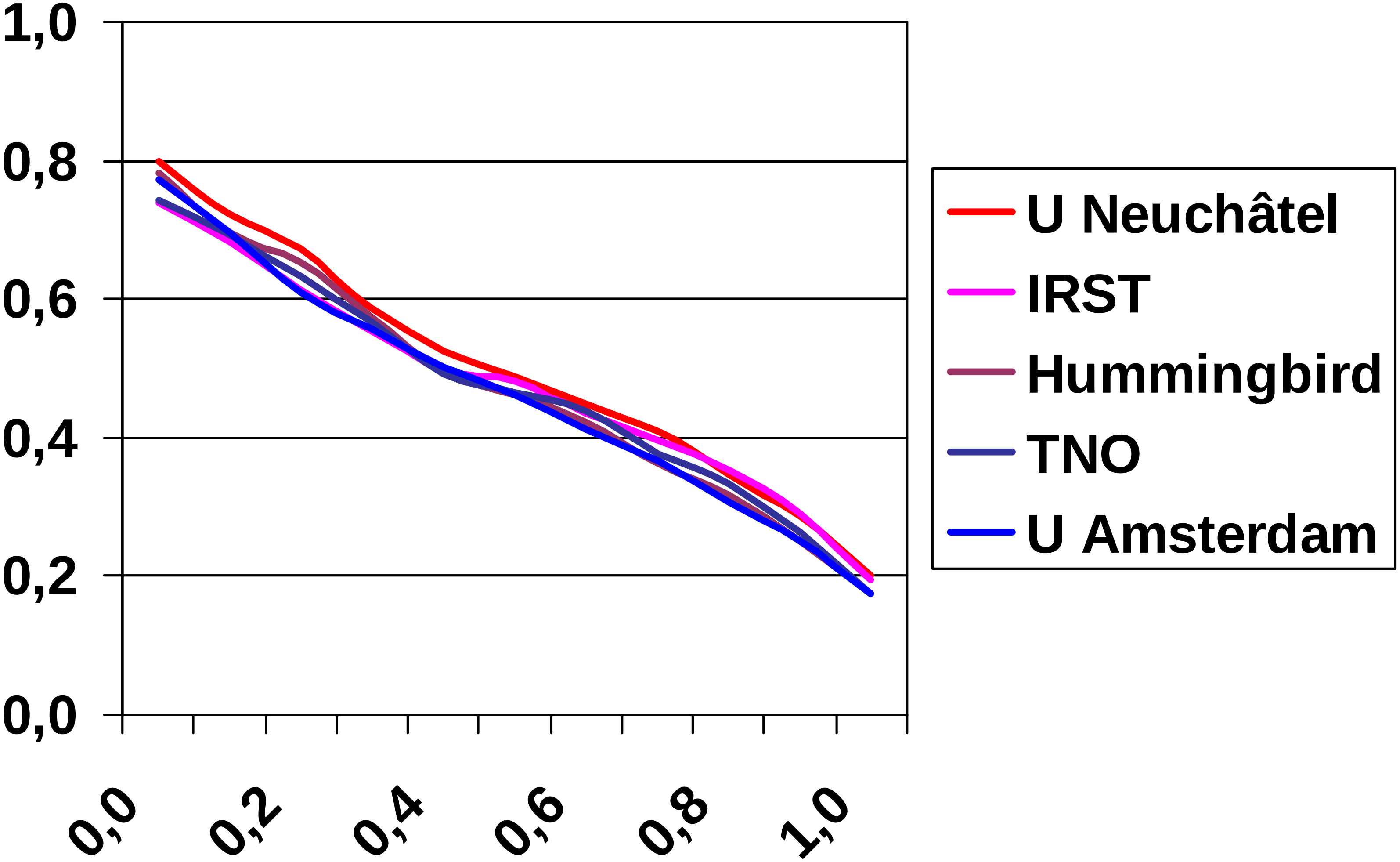


Eurospider

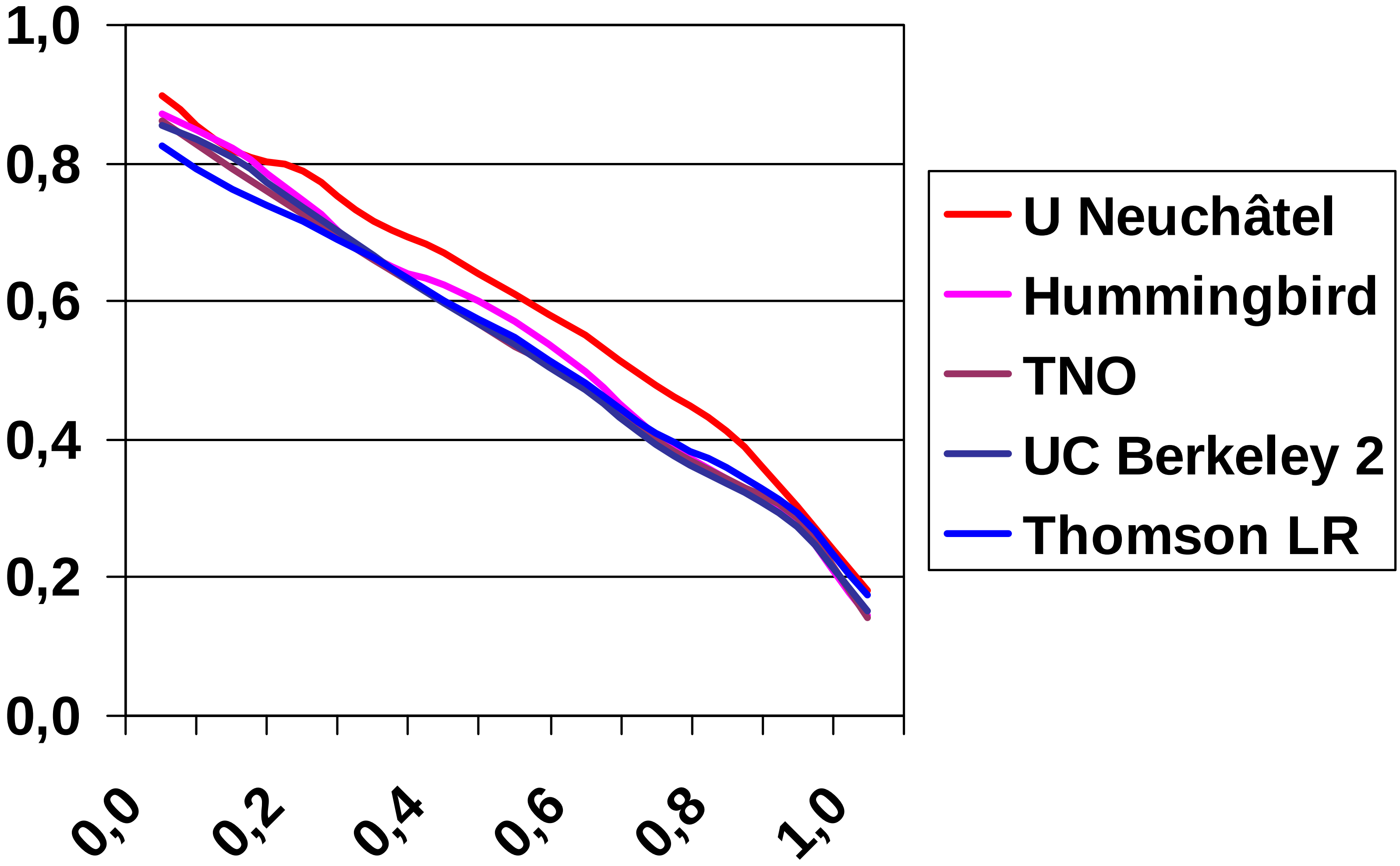
German


1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
Hummingbird
U Neuchâtel
Thomson LR
U Amsterdam
Eurospider

Italian


1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
U Neuchâtel
IRST
Hummingbird
TNO
U Amsterdam

Spanish


1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
U Neuchâtel
Hummingbird
TNO
UC Berkeley 2
Thomson LR