Retrieval Experiments in Unfamiliar Languages

Paul McNamee

Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel MD 20723-6099 USA
paul.mcnamee@jhuapl.edu

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

Unfamiliarity

- We can't remember whether it is Bulgarian or Hungarian that uses a Cyrillic character set
- No knowledge of morphological analyzers in Bulgarian or Hungarian
- No previous experience using Greek or Indonesian as query languages
- Truthfully, any non-English language is 'unfamiliar'
- Ignorance, Laziness, and Desperation, have driven JHU/APL to language neutral techniques

Language Neutrality – Why?

- **Simplicity**
  - No additional coding to support new languages
  - Avoid reliance of 3rd party software or data sources, that must be integrated
- **Scalability**
  - CLEF Newspaper collections: 12 languages
  - EuroGov corpus: 25 languages
- **Accuracy**
  - Competitive performance at CLEF in 12 languages without linguistic resources

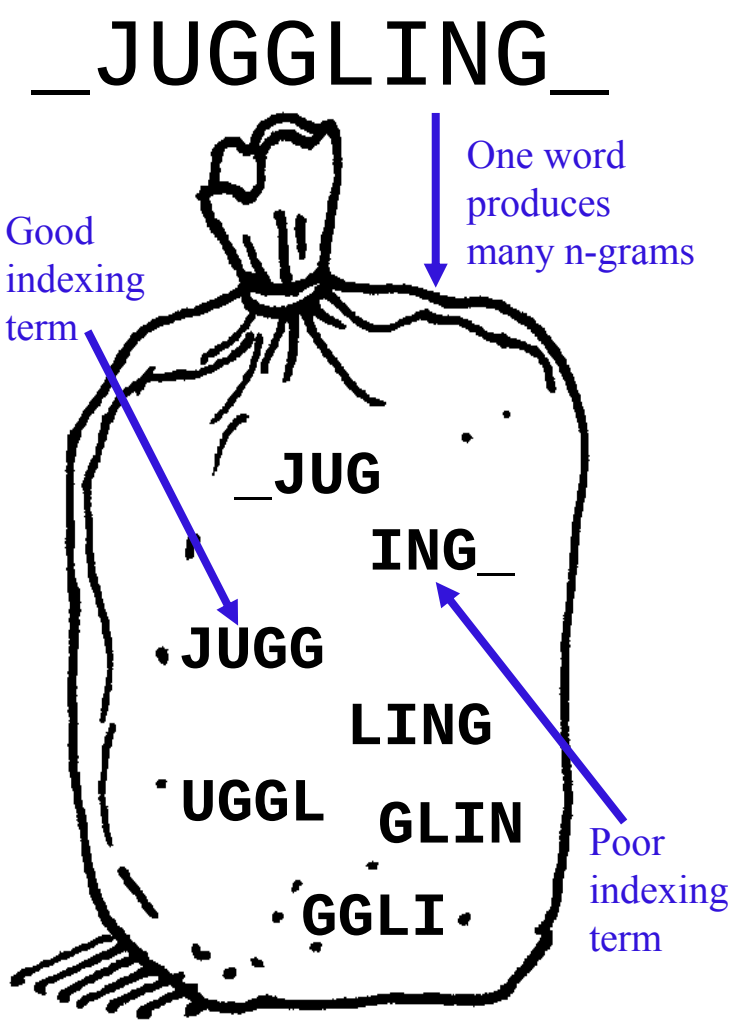
Talk Outline

- **Monolingual Experiments**
  - N-gram stemming (post hoc work)
  - N-grams vs. words
  - N-grams and Snowball stemmer
- **Bilingual Experiments**
  - With large parallel corpora (using n-grams)
  - Reliance on Web-based MT

Experimental Conditions (Monolingual)

- Used the JHU/APL HAIRCUT system
  - Java-based system described in 2004 article
    - P. McNamee and J. Mayfield, 'Character N-gram Tokenization for European Language Text Retrieval', *Information Retrieval* 7(1-2):73-97, 2004.
  - Default representation: space-delimited, accent-less words
- Statistical Language Model
  - Jelinek-Mercer smoothing – 1 smoothing parameter
- Uniform processing of each language
- Optional Blind Relevance Feedback
  - Expand query to 60 terms (25 top docs, 75 low docs)

N-Gram Tokenization



- Characterize text by overlapping sequences of *n* consecutive characters
- For alphabetic languages, *n* is typically 4 or 5
- N-grams are a language-neutral representation
- N-gram tokenization incurs both speed and disk usage penalties

"Every character begins an n-gram"

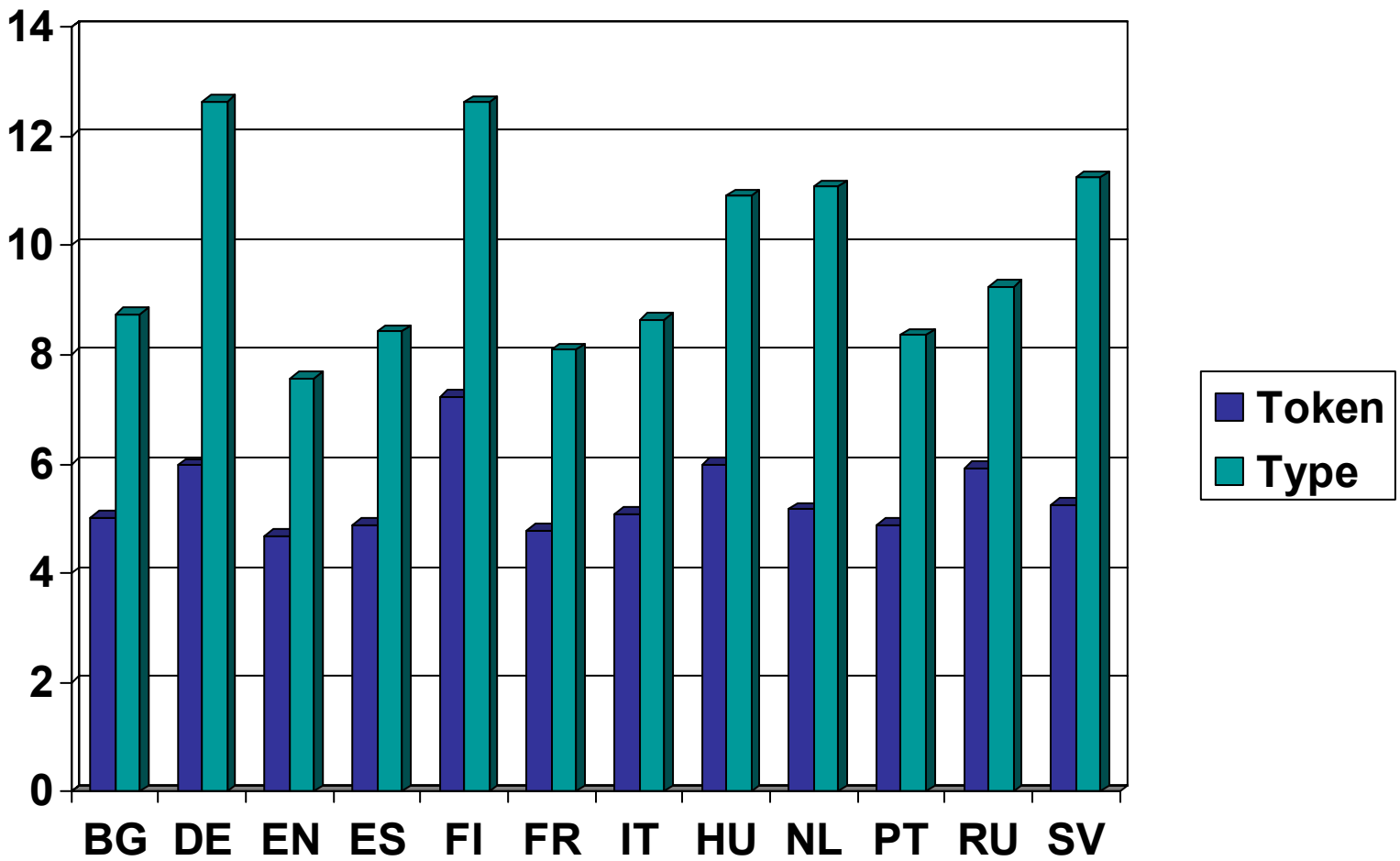
Query Processing With N-grams

	Mean Postings Length	Mean Response Time (secs)
7-grams	20.1	22.5
words	34.8	3.5
6-grams	44.2	30.6
5-grams	131.0	37.0
4-grams	572.1	37.2
3-grams	3762.5	14.5

CLEF 2002 Spanish Collection (1 GB)

- **A typical 3-gram will occur in many documents, but most 7-grams occur in few**
- **Longer n-grams have larger dictionaries and inverted files**
  - **But not longer response times**
- **N-gram querying can be 10 times slower!**
- **Disk usage is 3-4x**

Mean Word Length



N-gram Stemming

- **Traditional (rule-based) stemming attempts to remove the morphologically variable portion of words**
  - **Negative effects from over- and under-conflation**

Hungarian	Bulgarian
_hun (20547)	_bul (10222)
hung (4329)	bulg (963)
unga (1773)	ulga (1955)
ngar (1194)	lgar (1480)
gari (2477)	gari (2477)
aria (11036)	aria (11036)
rian (18485)	rian (18485)
ian_ (49777)	ian_ (49777)

Short n-grams covering affixes occur frequently - those around the morpheme tend to occur less often. This motivates the following approach:

(1) For each word choose the least frequently occurring character 4-gram (using a 4-gram index)

(2) Benefits of n-grams with run-time efficiency of stemming

Continues work in Mayfield and McNamee, 'Single N-gram Stemming', SIGIR 2003

Examples

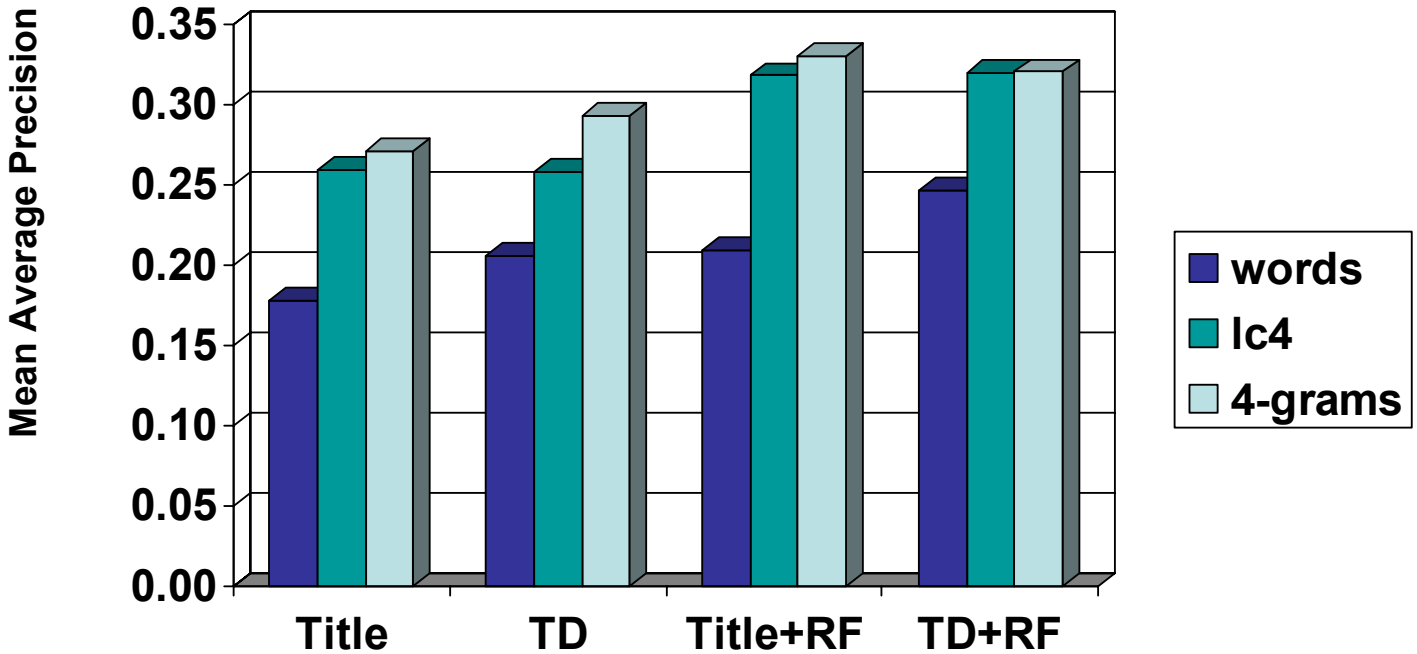
Lang.	Word	Snowball	LC4
English	juggle	juggl	jugg
English	juggles	juggl	jugg
English	juggler	juggler	jugg
English	juggled	juggl	jugg
English	juggling	juggl	jugg
English	juggernaut	juggernaut	rnau
English	warred	war	warr
English	warren	warren	warr
English	warrens	warren	rens
English	warrant	warrant	warr
English	warring	war	warr

Lang.	Word	Snowball	LC4
Swedish	kontroll	kontroll	ntro
Swedish	kontrollerar	kontroller	ntro
Swedish	kontrollerade	kontroller	ntro
Swedish	kontrolleras	kontroller	ntro
English	pantry	pantri	antr
English	tantrum	tantrum	antr
English	marinade	marinad	inad
English	marinated	marin	rina
English	marine	marin	rine
English	vegetation	veget	etat
English	vegetables	veget	etab

All approaches to conflation, including no conflation at all, make errors.

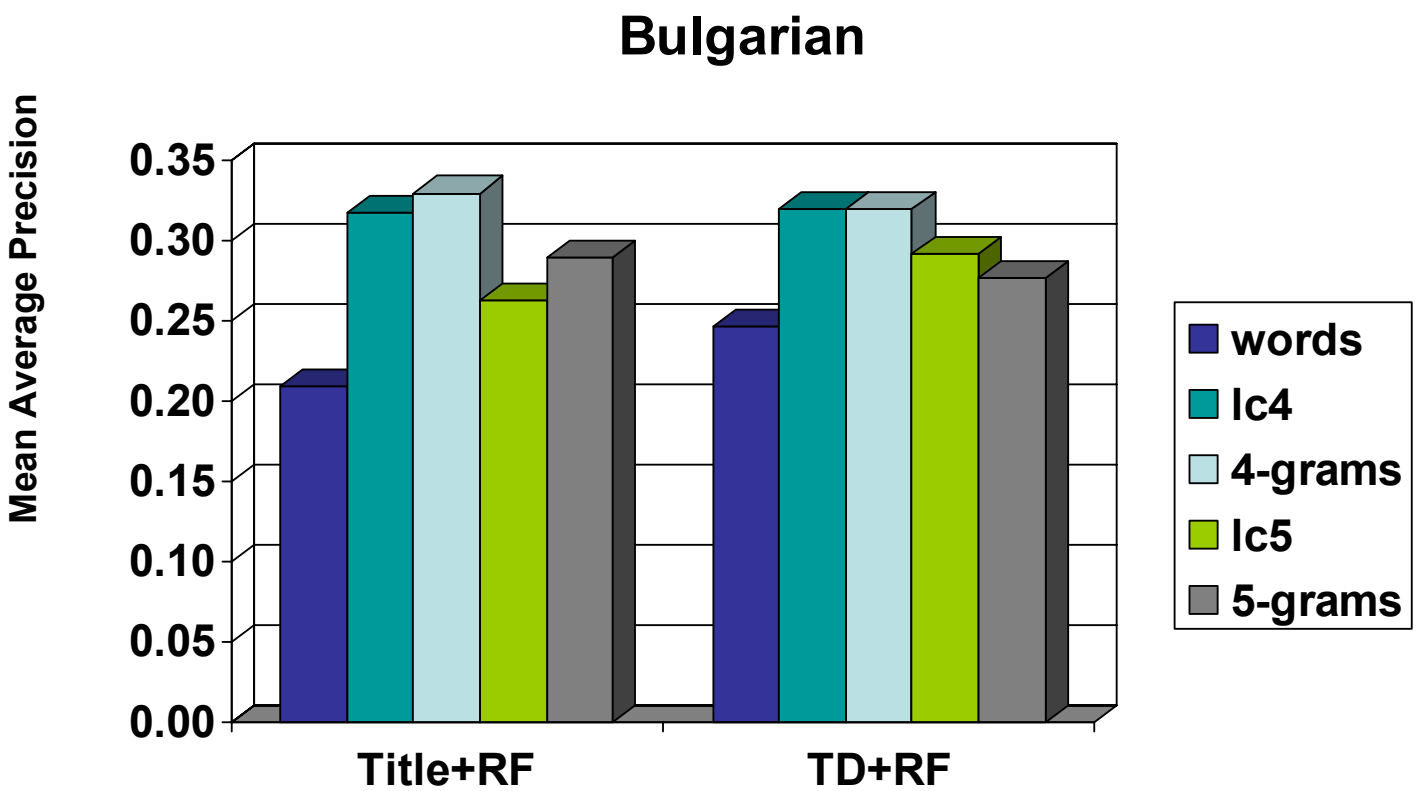
Tokenization Results (BG)

Bulgarian



- 4-grams dominate
- N-gram stemming 25-50% better than words, matches 4-grams
- PRF somewhat helpful with n-grams

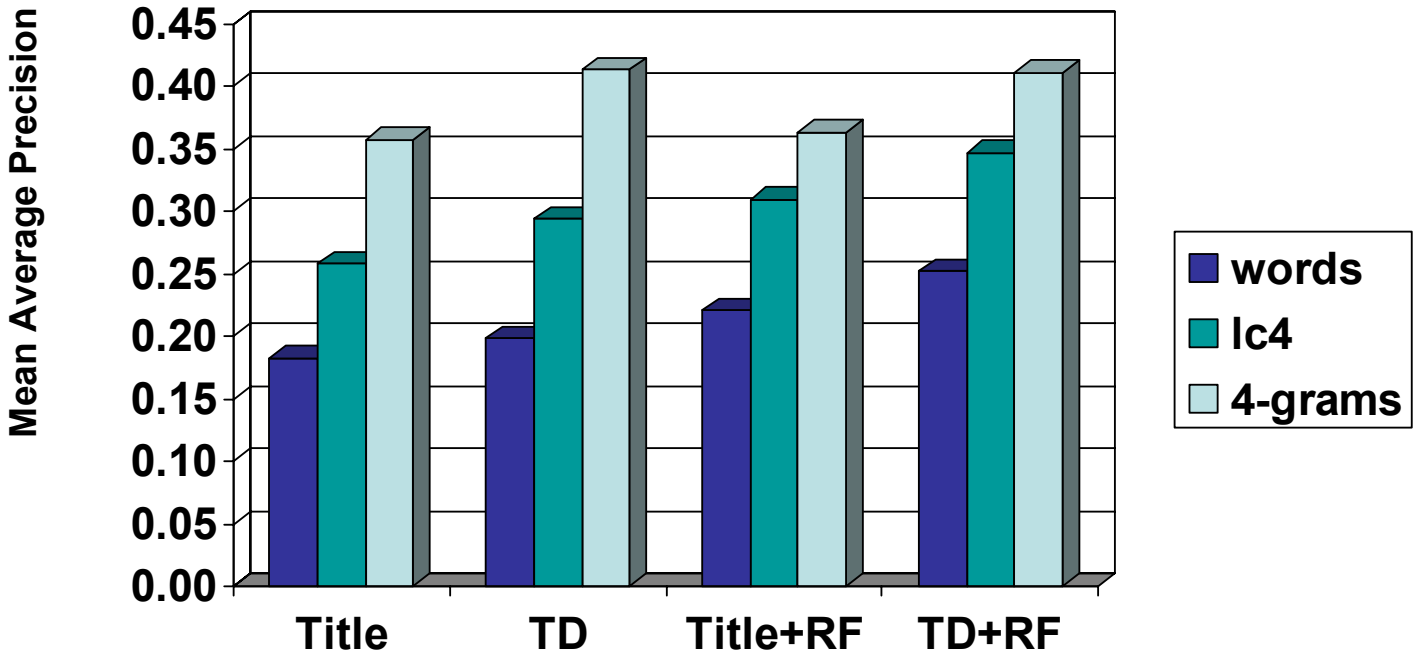
Tokenization Results (BG-2)



- 5-grams and 'Least-Common 5-gram stemming' not quite as effective as corresponding use of 4-grams

Tokenization Results (HU)

Hungarian



- 4-grams dominate
- N-gram stemming about 50% better than words
- PRF not helpful with 4-grams
- (not shown – results with 5-grams similar, but marginally worse than 4-grams)

Base Runs & Official Submissions

	BG	EN	FR	HU	PT	Remark
words	0.2464	0.4133	0.3749	0.2521	0.3453	
stems	--	0.4401	0.4378	--	--	
4-grams	0.3203	0.3692	0.3608	0.4112	0.3246	*aplmoxxd*
5-grams	0.2768	0.3873	0.3801	0.4056	0.3654	*aplmoxxe*
M(5+s)	--	0.4346	0.4214	--	--	*aplmoxxa*
M(4+s)	--	0.4222	0.4122	--	--	*aplmoxxb*
M(4+5)	0.3058	0.3898	0.3765	0.4063	0.3610	*aplmoxxc*

BG, HU: 4-grams decisively better (+30%, +60%)

EN, FR: Stems outperform n-grams

Bilingual Experiments

- **With Parallel Texts (3 runs)**
  - **'Enlarge' query using pre-translation expansion**
    - **CLEF source language collection used**
    - **Two runs merged (5-grams and stems)**
    - **60 words from top documents form expanded source language query**
  - **Revised query translated token-by-token**
    - **Tokens can be n-grams, words, or stems**
    - **500MB of aligned text from the Official Journal of the EU**
  - **Usually apply additional PRF**
- **Based on Web-MT (9 runs)**
  - **Translated query processed using words, stems, or n-grams, or a combination of the methods**

babelfish.altavista.com (GR - EN)

www.toggletext.com (IN - EN)

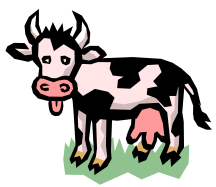
www.bultra.com (BG - EN)

www.tranexp.com (HU - EN)

Subword Translation

- Over the past couple of years we have experimented with statistically translated n-gram sequences
- Hope is to mitigate problems in dictionary-based CLIR
  - word lemmatization (variation in morphology)
  - out of vocabulary words, particularly names (few OOV n-grams)
  - multiword expressions (due to word-spanning n-grams)

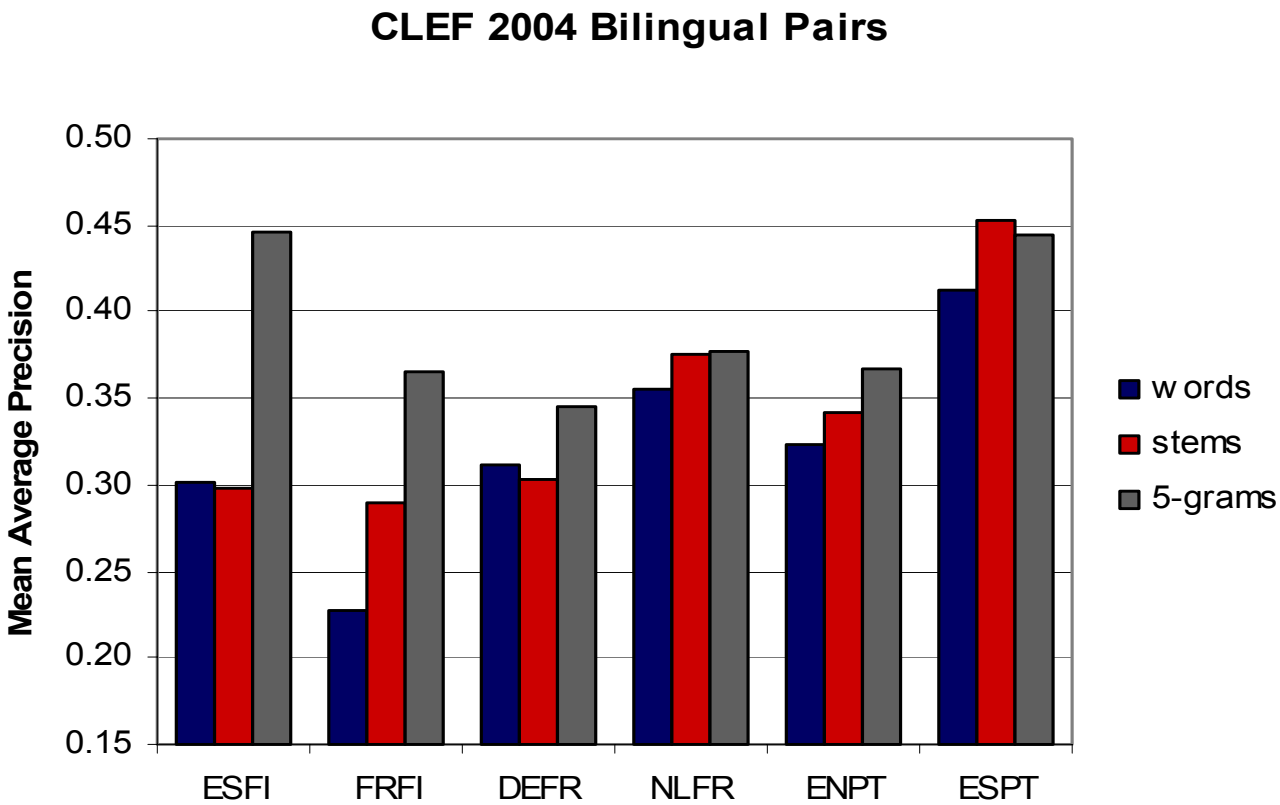
	German	Italian
word	milch	latte
stem	milch	latt
4-grams	milc ilch	latt latt
5-grams	_milc milch ilch_	_latt _latt latte



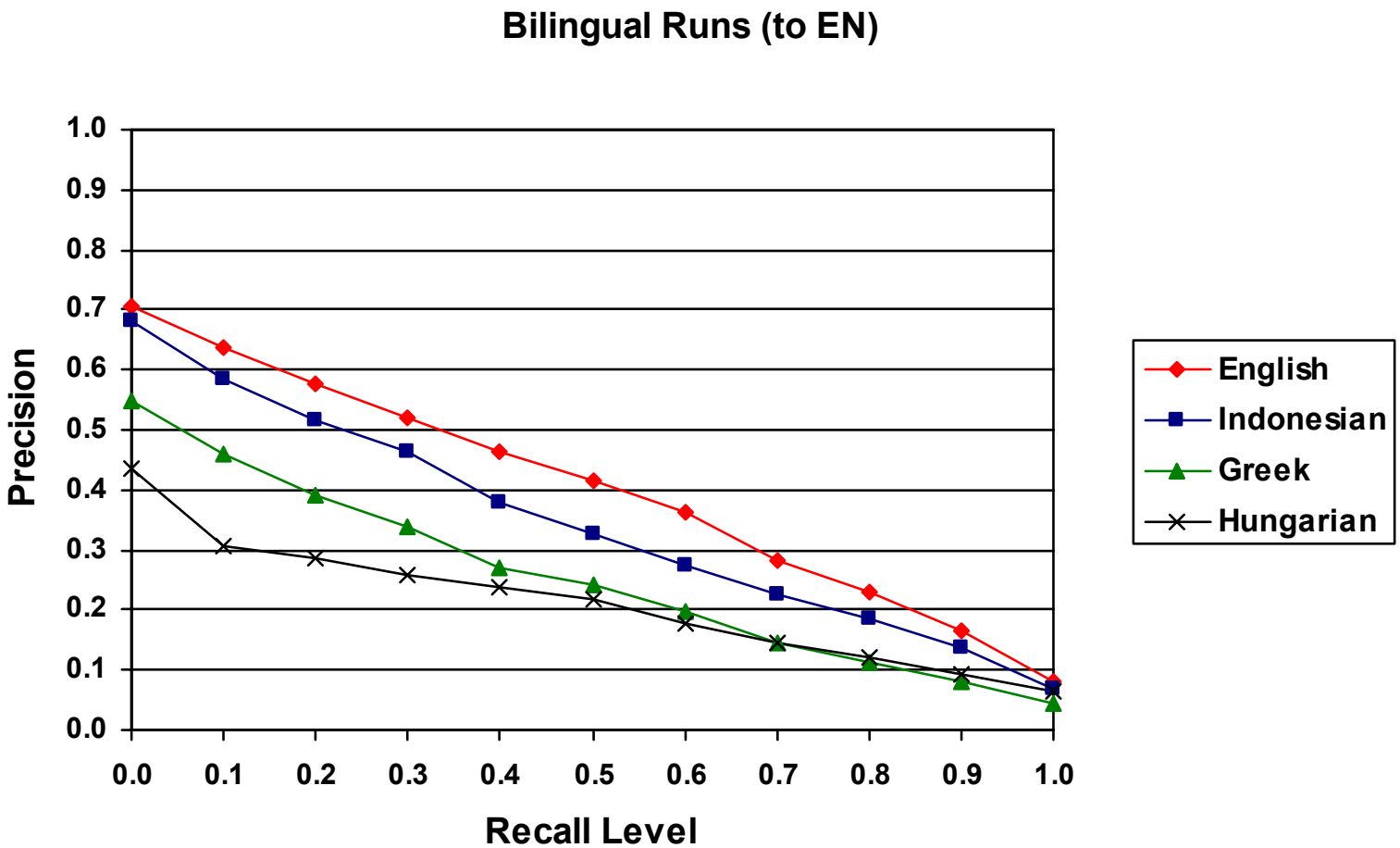
Official Results

		MAP	% mono
EN FR	TD 5-grams	0.3442	78.6%
EN PT	TD 5-grams	0.3130	85.4%
ES PT	TD 5-grams	0.3185	87.2%

Subword Translation Results



Results Using MT



Conclusions

- N-grams can be recommended for monolingual tasks
  - Especially true in languages with complex morphology and in which resources (*e.g.,* stemmers) not available
  - 4-grams quite effective (+60% advantage over HU words)
- Query-time penalty of n-grams can be reduced, with some concomitant loss of effectiveness
  - 'N-gram stemming' yields 30-50% improvements over plain words in Bulgarian and Hungarian
- Bilingual retrieval with n-grams is also attractive
  - Subword translation using aligned corpora is effective
    - Caveat: must have parallel data
  - N-grams can also be used with MT


Sharon McNamee
September 2005, Louisiana
OPERATION
LIFELINE