
# N-grams and Morpheme Analysis in IR

**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

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# Outline

- **Character N-grams in IR**
  - Confusing History
- **Empirical Studies**
  - Comparision with plain words
  - Problems with Efficiency
  - Synthetic Morphology (N-gram stemming)
  - MorphoChallenge 2007
- **Summary**

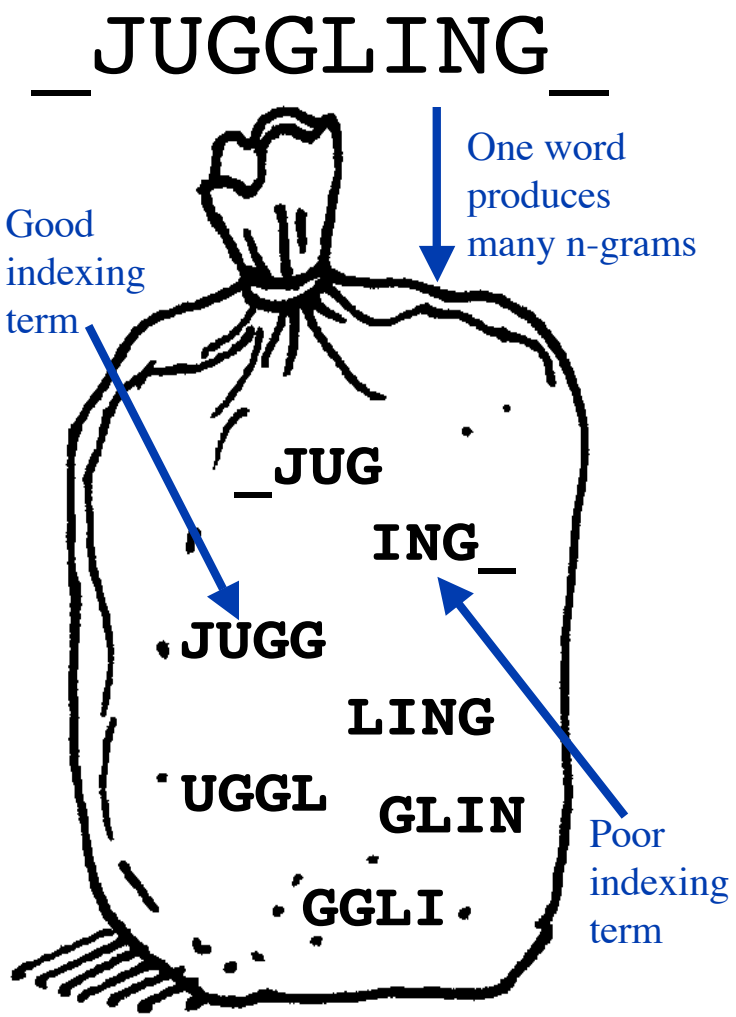




JOHNS HOPKINS
UNIVERSITY
Applied Physics Laboratory

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# N-Gram Tokenization




- Characterize text by overlapping sequences of *n* consecutive characters
- In 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"

# Against: Damashek (1995)

- Marc Damashek developed an IR system based on n-grams
  - *'Gauging Similarity with n-Grams: Language Independent Categorization of Text'*, Science, vol. 267, 10 Feb 1995
  - He described his system's performance at TREC-3 as:
    - "on a par with some of the best existing retrieval systems."

- The article elicited strong reaction
  - TREC Program Committee objected stating his system was ranked 22/23 and 19/21 on two tasks
  - IR luminary Gerald Salton wrote a response
    - "decomposition of running texts into overlapping n-grams ... is too rough and ambiguous to be usable for most purposes."
    - "for more demanding tasks, such as information retrieval, the n-gram analysis can lead to disaster"
    - "decomposition of text words such as HOWL into HOW and OWL raises the ambiguity of the text representation and lowers retrieval effectiveness"

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# Pro: Asian Languages (1999)

- *Information Processing and Management* 35(4) was devoted to IR in Asian Languages
  - Many Asian languages lack explicit word boundaries
- Korean
  - Lee et al., KRIST Collection (13K docs)
    - 2-grams outperform words, decompounding cited
- Chinese
  - Nie and Ren, TREC 5/6 Chinese Collection (165K docs)
    - 2-grams (0.4161 avg. prec.) comparable to words (0.4300)
    - Combination of both is best (0.4796)
- Japanese
  - Ogawa and Matsuda, BMIR-J2 (5K docs)
    - M-grams (unigrams and bigrams) comparable to words

# Against: "A Basic Novice Solution"

WHAT'S NEXT

## ...m Uzbek to Klingon, the Machine Cracks the Code

...R JOHN FARAH

...99, at a workshop on ... translation at Johns ...ty, Kevin Knight ...an advertisement to ...search team he was ...f the ad was a picture ...archment covered in ...To most people, this ...," the ad announced. ...e broken."

...duct yet to be created ...ur in a new bunch of ..., alongside a picture ...e think you'll be sur-

...eant to be a motiva- ...the field of statistical ...s all but dead. In the ...ssed since that work- ...ad of machine trans- ...University of South- ...lation Sciences Insti- ...how prophetic the ad ...are," he said. "It's no

... translation — in ...tially learn new lan- ... instead of being ... by bilingual human ...taken off. The new ...ntists to develop ma- ...ms for a wide num- ...es at a pace that ex- ...possible.

...rs said the progress ...ical machine transla- ...ssed that of the tradi-tional machine translation programs used by Web sites like Yahoo and BabelFish. In the past, such programs were able to compile extensive databanks of foreign languages that allowed them to outperform statistics-based systems.

Traditional machine translation relies on painstaking efforts by bilingual programmers to enter the vast wealth of information on vocabulary and syntax that the computer needs to translate one language into another. But in the early 1990's, a team of researchers at I.B.M. devised another way to do things: feeding a computer an English text and its translation in a different language. The computer then uses statistical analysis to "learn" the second language.

Compare two simple phrases in Arabic: "rajl kabir" and "rajl tawil." If a computer knows that the first phrase means "big man," and the second means "tall man," the machine can compare the two and deduce that rajl means "man," while kabir and tawil mean "big" and "tall," respectively. Phrases like these, called "N-grams" (with N representing the number of terms in a given phrase) are the basic building blocks of statistical machine translation.

Although in one sense it was more economical, this kind of machine translation was also much more complex, requiring powerful computers and software that did not exist for most of the 90's. The Johns Hopkins workshop changed all that, yielding a software application package, Egypt/Giza, that made statistical translation accessible to researchers across the country.

"We wanted to jump-start a vibrant field," Dr. Knight said. "There was no software or data to play with."


Mary Ann Smith

Today researchers are racing to improve the quality and accuracy of the translations. The final translations generally give an average reader a solid understanding of the original meaning but are far from grammatically correct. While not perfect, statistics-based technology is also allowing scientists to crack scores of languages in a fraction of the time, and at a fraction of the cost, that traditional methods involved.

A team of computer scientists at Johns Hopkins led by David Yarowsky is developing machine translations of such languages as Uzbek, Bengali, Nepali ... "Star Trek."

"If we can learn how t... Klingon into English, then ... guages are easy by comp... "All our techniques require ... two languages. For exam...

Language Institute translated 'Hamlet' and the Bible into Klingon, and our programs can automatically learn a basic Klingon-English MT system from that."

Dr. Yarowsky said he hoped to have working translation systems for as many as 100 languages within five years. Although the grammatical structures of languages like Chinese and Arabic make them hard to analyze statistically, he said, it will only be a matter of time before such hurdles are overcome. "At some point, we start encountering the same problems over and over," he said.

In addition to the release of Egypt/Giza in ...

**Armed with an English text and a translation, a computer uses statistical analysis to 'learn' the second tongue.**

1999, the spread of the Internet has led to an explosion of translated texts in far-flung languages, greatly aiding the team's research. Researchers have also benefited from a much faster means of evaluating the outcome of translation experiments: a computerized technique developed by I.B.M. enables researchers to test 10 to 100 new approaches for cracking languages each day.

provides scientists with a fast, objective measurement that they can use to note improvement and saves them from having to review every unsuccessful experiment.

"Before Bleu, it was really a bad state of affairs," said Alex Fraser, a doctoral student at U.S.C. "You look at broken couplets of English for a long time, and eventually you start to accept it more and more."

Despite the progress being made in statistical machine translation, some researchers remain skeptical, preferring to focus their efforts on language-specific translation techniques. Ophir Frieder, a professor of computer science at the Illinois Institute of Technology, is working on a search system exclusive to Arabic text.

"Yes, N-grams work on any language, but as a search technique they work poorly on every language," he said. "It's a basic novice solution."

Dr. Knight acknowledges that statistical machine translation is far from perfect. In its latest efforts, his team has sought to combine the statistical and traditional approaches to achieve ...imum accuracy and to produce trans... that the average computer user can ...stand. The best machine translation ...ms today, while capable of yield... ...age's general meaning, are better known for their muddled syntax than their accuracy. By applying the principles of statistical translation to

**"Yes, N-grams work on any language, but as a search technique they work poorly on every language," he said. "It's a basic novice solution."**

**-quote attributed to an IR researcher in the New York Times on 31 July 2003**

# The Truth is Out There...

## What should we conclude?

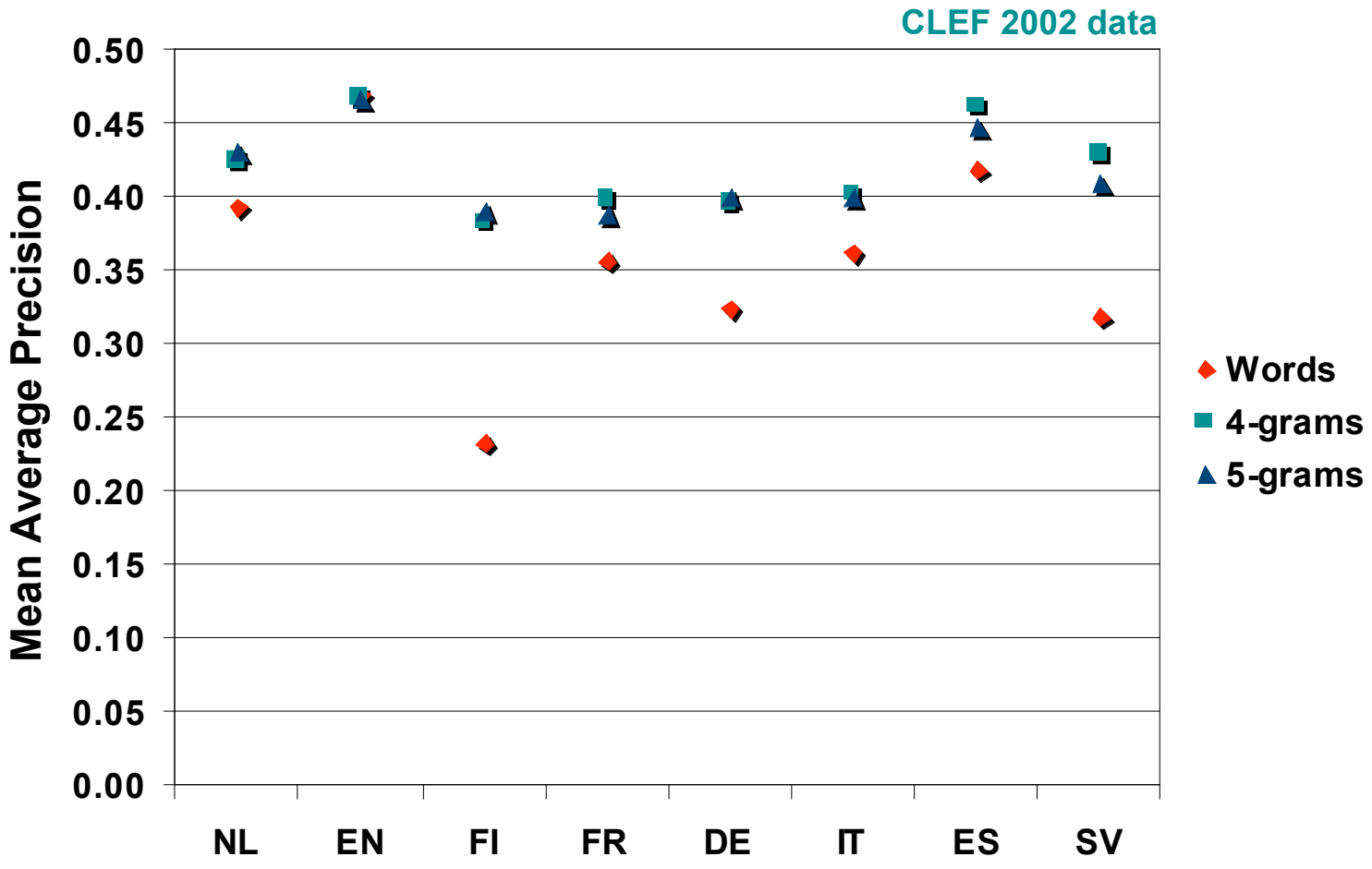
1. N-grams are not effective
2. N-grams are effective, but only in Asian Languages
3. Some IR Researchers do not like n-grams
4. Something else?

# HAIRCUT

- The *Hopkins Automatic Information Retriever for Combing Unstructured Text* (HAIRCUT)
  - Written in Java for portability and ease of implementation
  - Language-neutral philosophy
  - Language Model similarity measure
    - Ponte & Croft, 'A Language Modeling Approach to Information Retrieval,' SIGIR-98
    - Miller, Leek, and Schwartz, 'A Hidden Markov Model Inform Retrieval System', SIGIR-99.
  - Flexible tokenization schemes (e.g., n-grams)
  - Supports massive lexicons

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# Words vs. N-grams




From McNamee and Mayfield, 'Character N-gram Tokenization for European Language Text Retrieval.' *Information Retrieval* 7(1-2):73-97, 2004.

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# CLEF 2003 Monolingual Base Runs

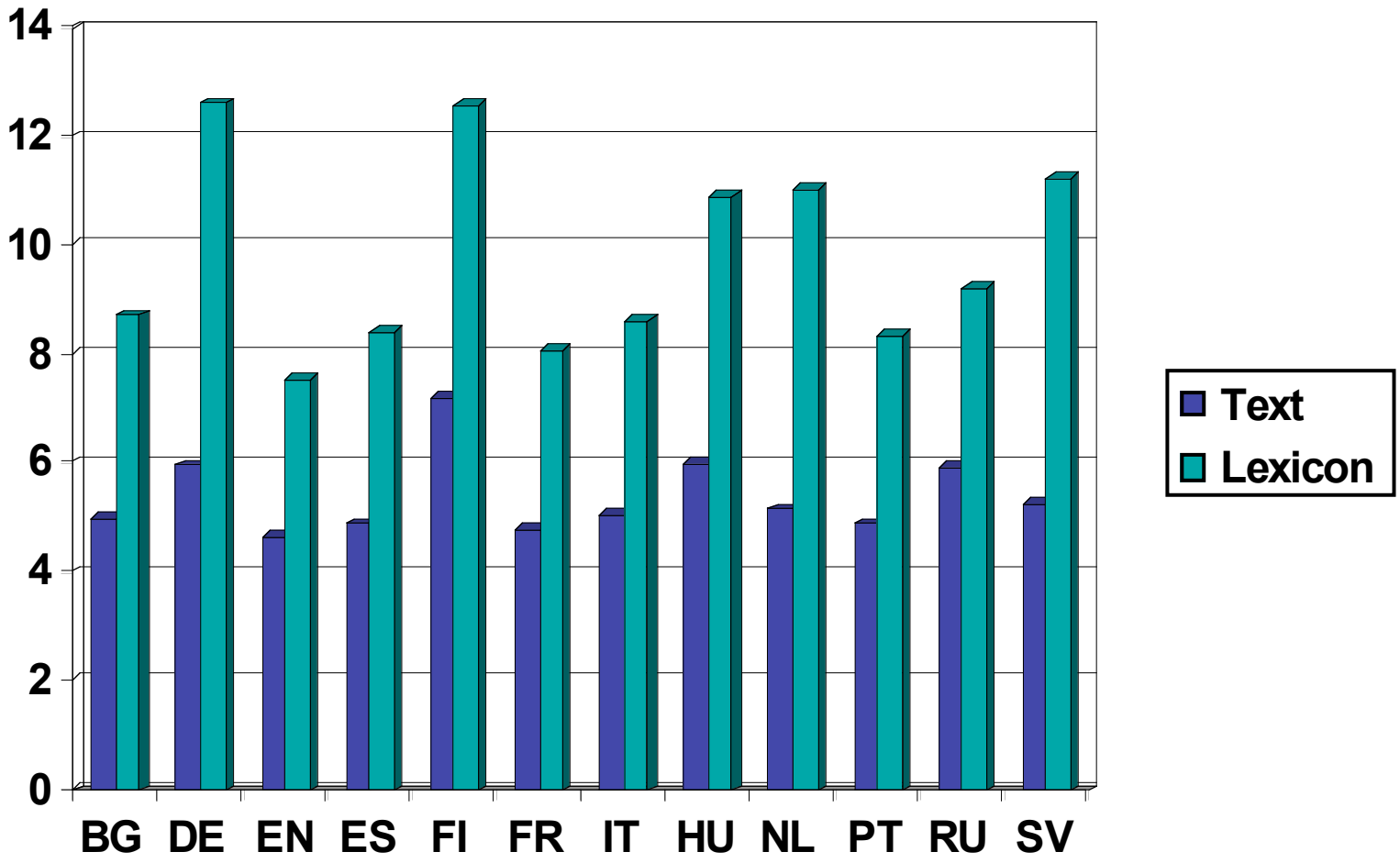
| | # topics | words | stems | 4-grams | 5-grams | Fusion |
|---|---|---|---|---|---|---|
| DE | 56 | 0.4175 | 0.4604 | 0.5056 | 0.4869 | 0.5210 |
| EN | 54 | 0.4988 | 0.4679 | 0.4692 | 0.4610 | 0.5040 |
| ES | 57 | 0.4773 | 0.5277 | 0.5011 | 0.4695 | 0.5311 |
| FI | 45 | 0.3355 | 0.4357 | 0.5396 | 0.5498 | 0.5571 |
| FR | 52 | 0.4590 | 0.4780 | 0.5244 | 0.4895 | 0.5415 |
| IT | 51 | 0.4856 | 0.5053 | 0.4313 | 0.4568 | 0.4784 |
| NL | 56 | 0.4615 | 0.4594 | 0.4974 | 0.4618 | 0.5088 |
| RU | 28 | 0.2550 | 0.2550* | 0.3276 | 0.3271 | 0.3728 |
| SV | 53 | 0.3189 | 0.3698 | 0.4163 | 0.4137 | 0.4358 |

**Single best monolingual technique: 4-grams**

**Fusion helpful, except in Italian**

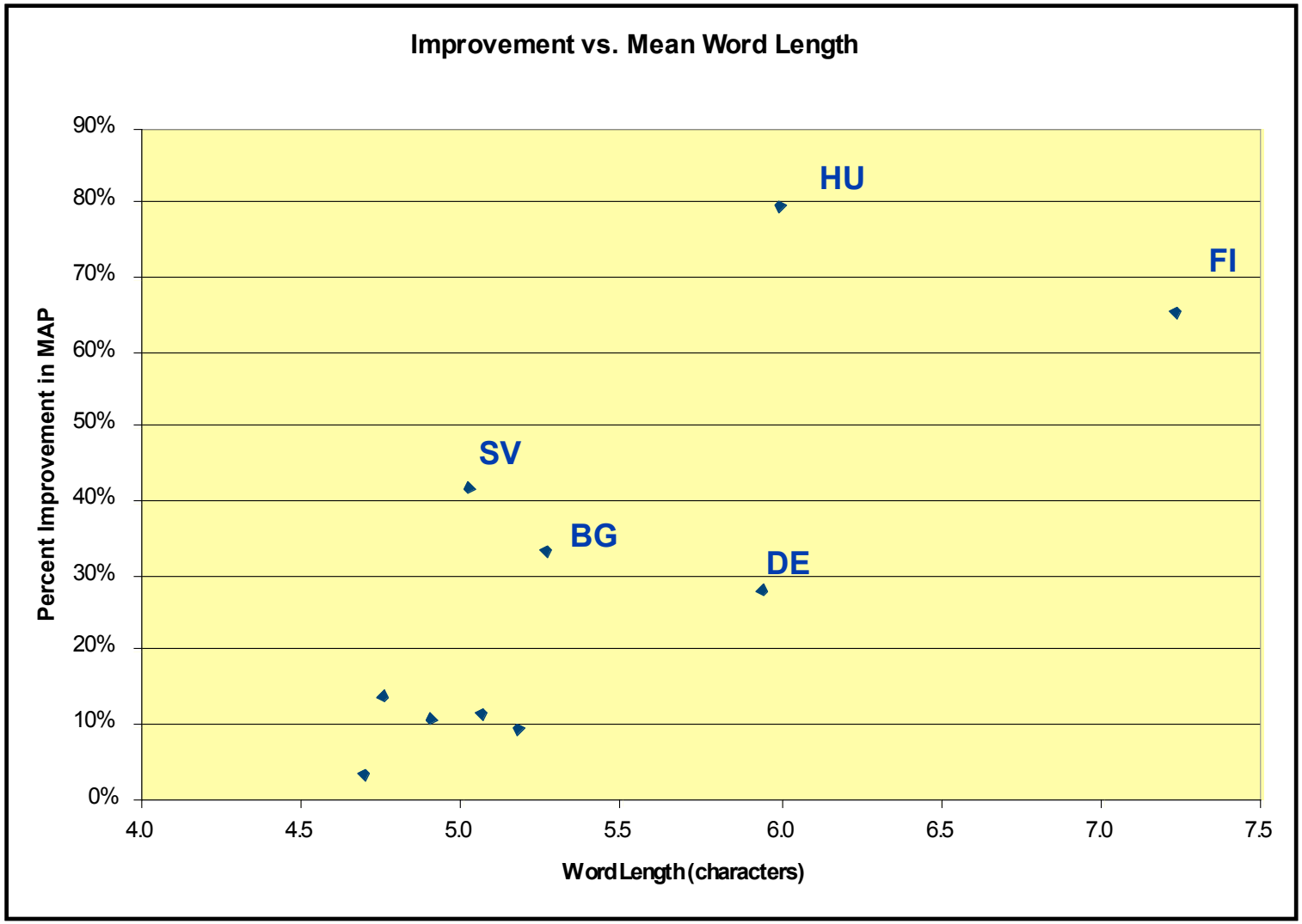
RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# Mean Word Length



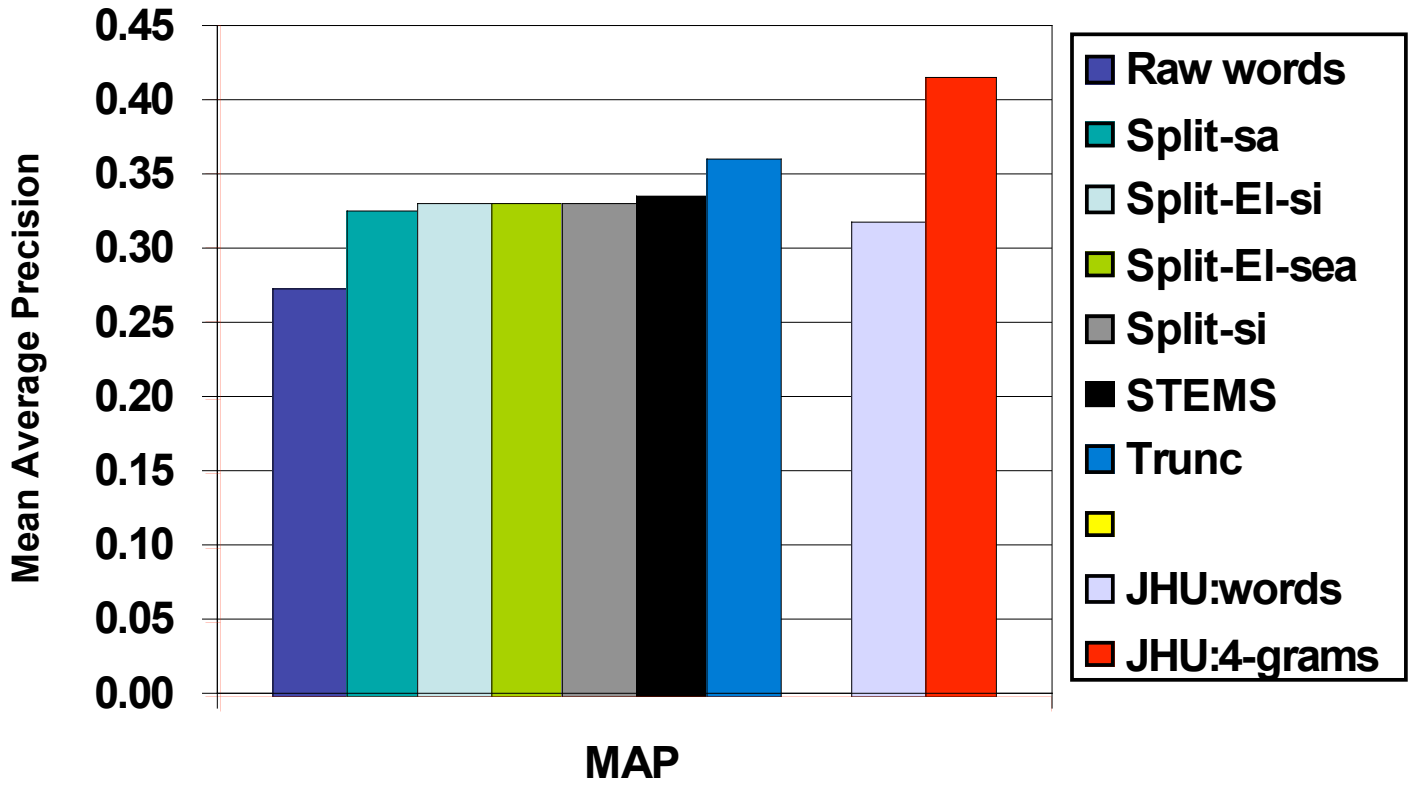
RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# N-grams vs. Words



RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

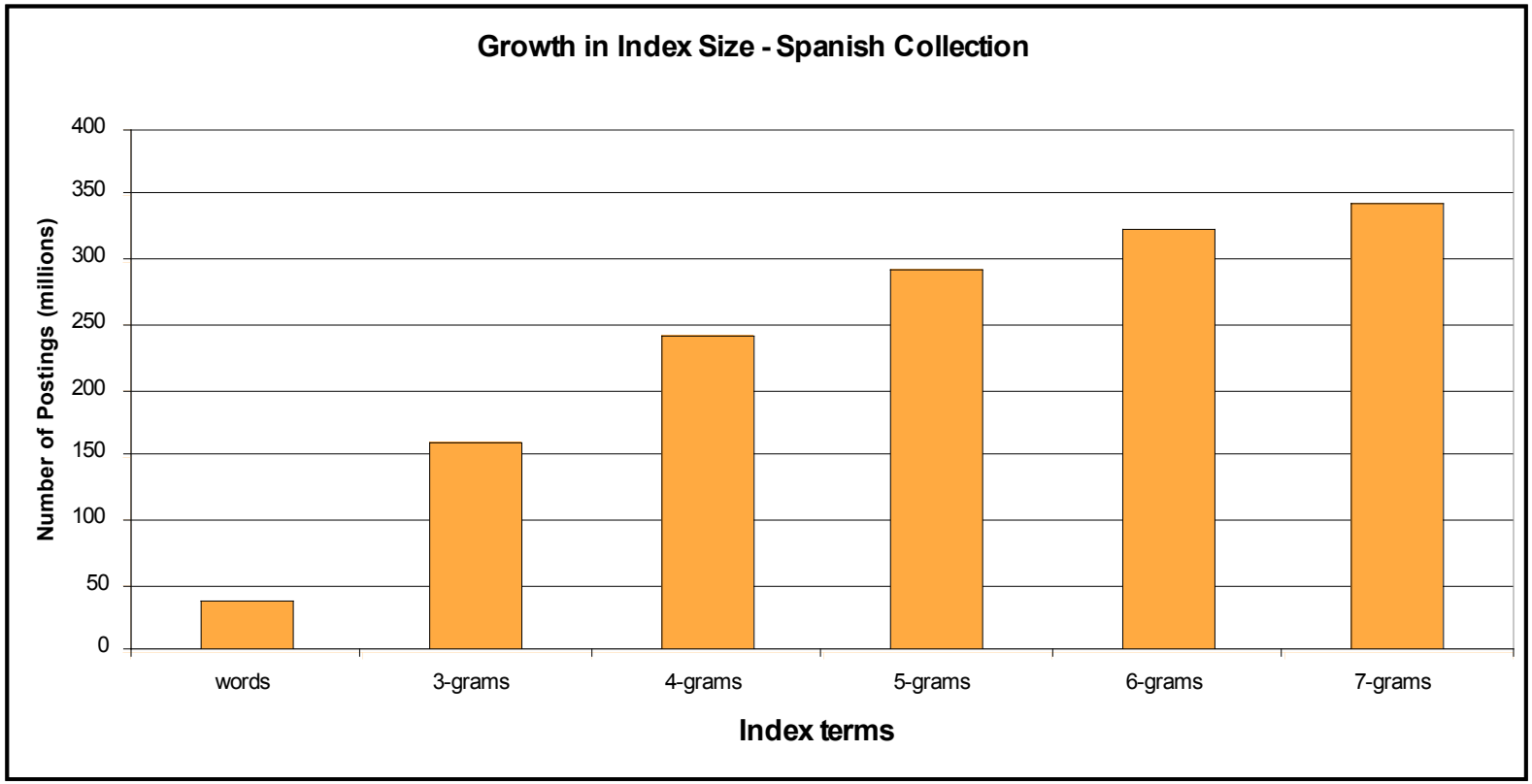
# Swedish Retrieval (CLEF 2003)




**Ahlgren and Kekalainen, 'Swedish Full Text Retrieval: Effectiveness of different combinations of indexing strategies with query terms'. Information Retrieval 9(6), Dec. 2006.**

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# N-gram Indexing: Size Matters



RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# 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**

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# 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**

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# 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 |

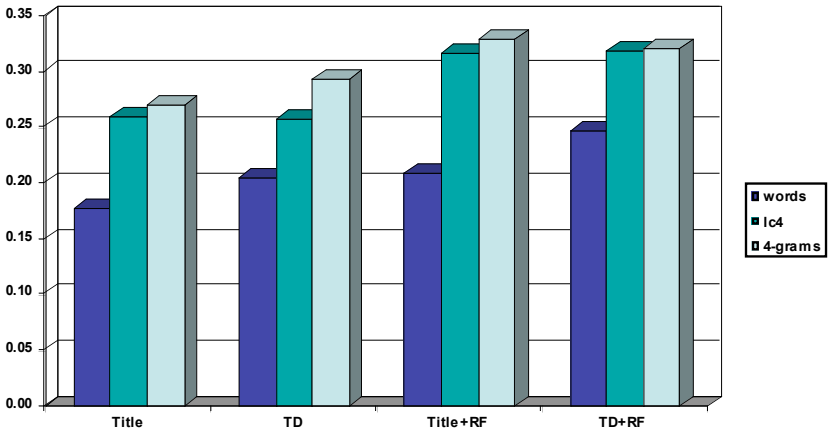
| Lang. | Word | Snowball | LC4 |
|---|---|---|---|
| Swedish | kontroll | kontroll | ntro |
| Swedish | kontrollerar | kontroller | ntro |
| Swedish | kontrollerade | kontroller | ntro |
| Swedish | kontrolleras | kontroller | ntro |
| English | pantry | pantri | antr |
| English | tantrum | tantrum | 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.**

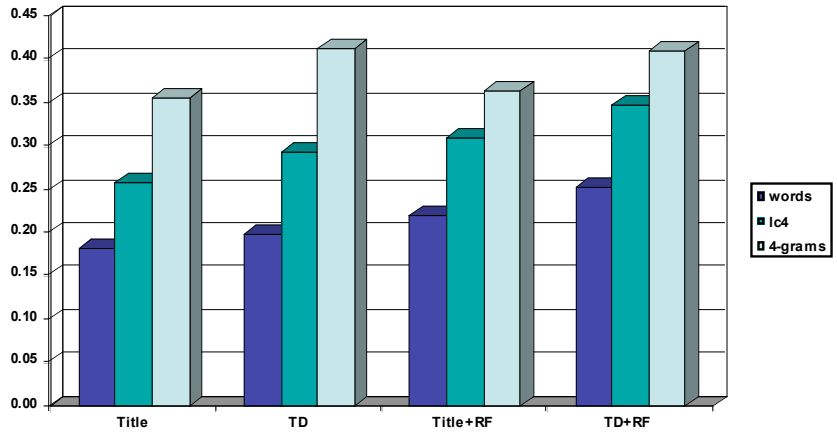
RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# N-gram Effectiveness

**Bulgarian**



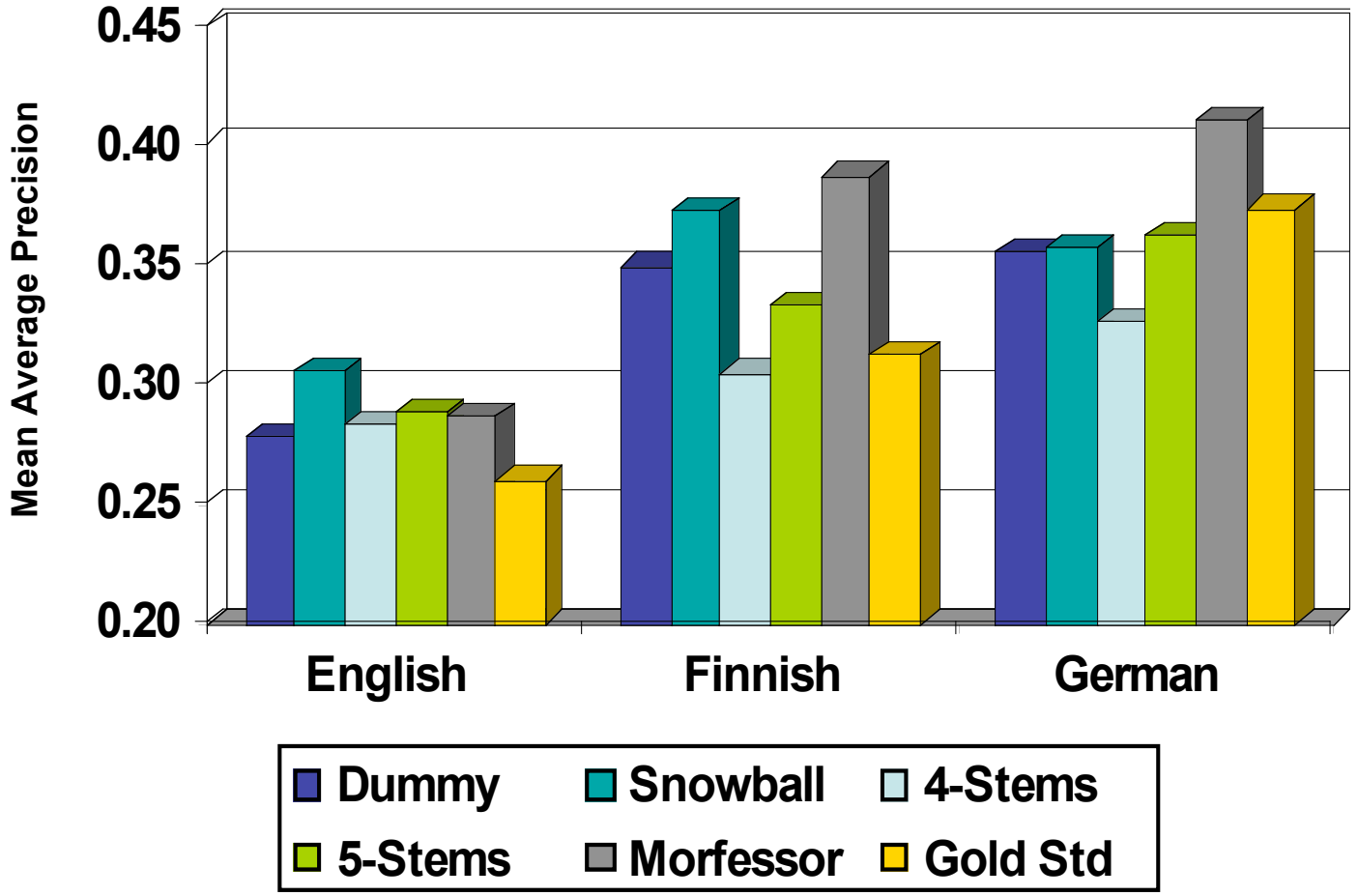

**Hungarian**




- **4-grams dominate words**
  - **25-50% advantage in Bulgarian**
  - **Improvements even larger in Hungarian**
- **4-gram stemming also dominates words**
- **Advantage consistent with and w/o blind feedback**

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# MorphoChallenge Task 2




**Withnew/TFIDF condition. 5-Stems beat 4-Stems. Morfessor is the clear winner.**

JOHNS HOPKINS
UNIVERSITY
Applied Physics Laboratory

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# Damashek revisited

- In 1995 no empirical evidence existed to support adequacy or supremacy of n-grams for IR
- N-grams appear less advantageous for English
- N-grams are conflationary
  - Salton was right (and wrong)
    - HOWL -> HOW, OWL
  - Longer and overlapping n-grams are more discriminating
    - HOWL, HOWLING, HOWLED, HOWLS share _HOW, HOWL

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

# Summary

- **N-grams very effective in European languages**
  - **As good or better than words and Snowball-produced stems**
  - **N=4 or N=5 both highly effective across CLEF languages**
  - **Numerous advantages, albeit performance issues**
    - **Don't need sentence splitter, tokenizer, stopword list, lexicon, thesaurus, stemmer**
    - **Simplicity for dealing with many languages**
- **Frequency-based n-gram stemming works**
  - **Benefit of n-grams or stemming, without any performance penalty**
  - **Available in all languages without customization**
  - **In compounding languages, a single n-gram may not be enough**