1st Workshop on Digital Information Management
30-31 March 2011, Corfu, Greece

"Failed Queries: a Morpho-Syntactic Analysis Based on Transaction Log Files"

Anna Mastora[1], Maria Monopoli[2] and Sarantos Kapidakis[1]

[1]Laboratory on Digital Libraries & Electronic Publishing, Department of Archives & Library Sciences, Ionian University, 72 Ioannou Theotoki str., 49100, Corfu, Greece

[2]Library Section, Economic Research Department, Bank of Greece, 21 El. Venizelos ave., 10250, Athens, Greece

{mastora, sarantos}@ionio.gr, mmonopoli@bankofgreece.gr

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

Presentation outline

- Introduction
- Aims and Objectives
- Related Research
- Definitions and Methodology
- Results
- Conclusions
- Future Work

Aims and Objectives

- The aim of this study is to elaborate on the procedure needed in order to analyze morpho-syntactically the typing-error queries submitted during the search process
- The objectives of the study are twofold
  - Explore the extent and types of failed queries due to typing errors
  - Explore the feasibility of their morpho-syntactic analysis

Related research (i)

- Information Retrieval techniques do not work effectively at all times
- "Not working effectively" includes:
  - Not retrieving relevant documents (low Recall), or,
  - Retrieving non relevant documents (low Precision)
- Part of studying what is not retrieved is the analysis of "Failed queries" or "Failure analysis"
- Significant interest has been expressed on failed queries as the outcome of subject searching

Related research (ii)

- Natural Language Processing is essential, especially in highly inflectional languages, such as Greek
- Part of Speech (PoS) tagging and morpho-syntactic analysis of words are valuable tools when mechanisms are applied for word disambiguation (either morpho-syntactic or word-sense driven)

Definitions and Limitations (i)

- What constitutes a "failed query"?
  - It depends. Some approaches consider:
    - Precision and Recall, applying retrieval effectiveness measures
    - User satisfaction, applying users' criteria to measure if a query was failed
    - Transaction Log files analysis (with or without relevance judgments)
    - Critical incident technique, using direct observation of human behavior

Definitions and Limitations (ii)

- **Failed query: a query with zero hits**
  - A bit arbitrary (yet practical) since
    - zero hits cannot solidly define a query as failed
    - not all non-zero hits queries can be defined as successful
  - For the purpose of this study this is a rather safe definition since
    - the database was known to include relevant items
    - the information need was relevant to the context of the database
    - we used transaction log files without relevance judgments
  - We analyze the queries morpho-syntactically, i.e. as "bag of words", detached from any semantic designation (well… most of the times…)

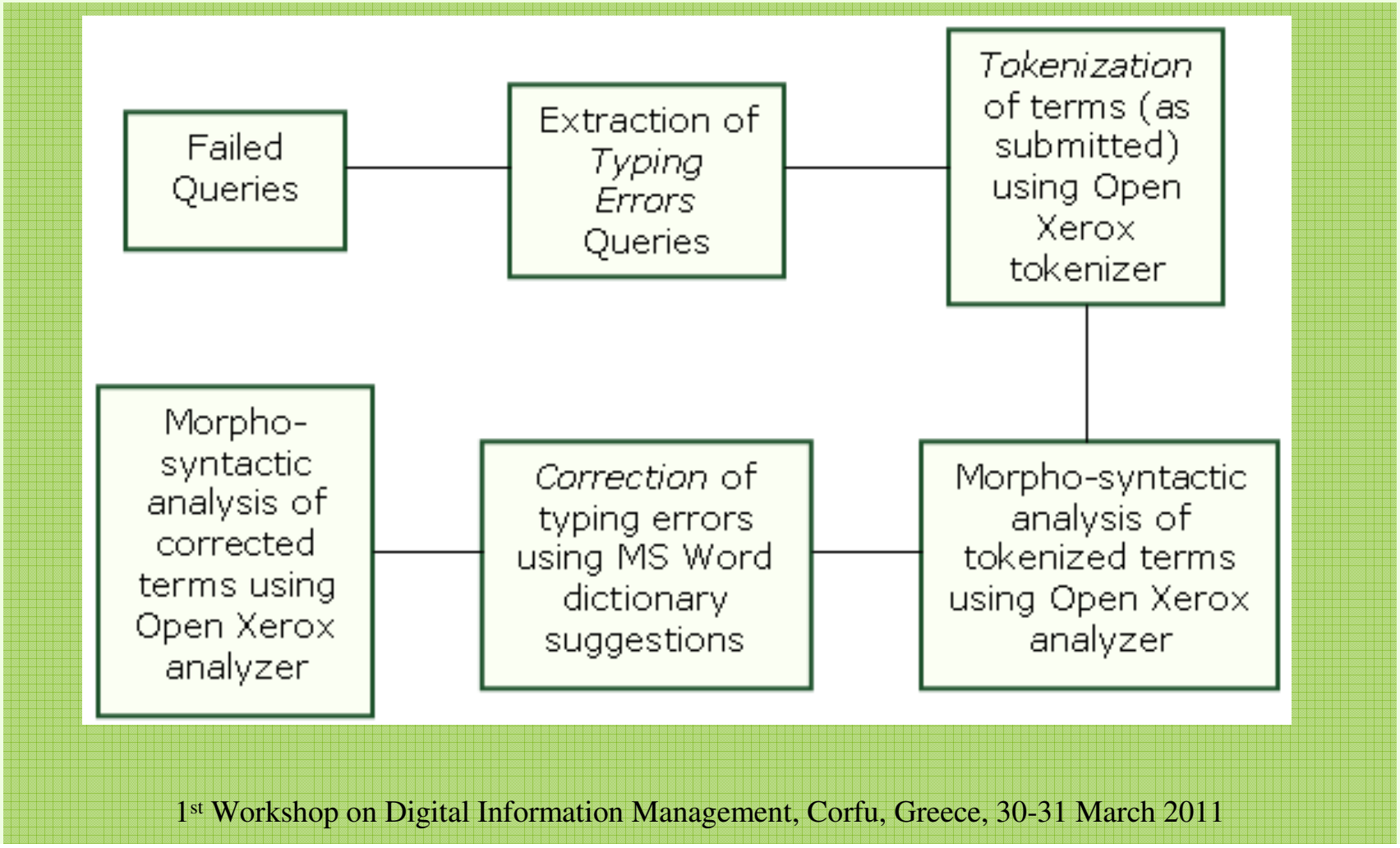
Definitions and Limitations (iii)

- **Morpho-syntactic analysis:** cognitive process that constitutes an intermediate layer between morphological and syntactic analysis and aims to assign unambiguous morpho-syntactic information to words of texts
- **Morpho-syntactic information:** the morphological origin and the morpho-syntactic properties of a word (e.g. the word *ανθρώπου* is the genitive singular form of the masculine noun [*άνθρωπος*])
- **Inflectional languages:** with a high morpheme-per-word ratio
- **Morpheme:** the smallest meaningful linguistic unit

Methodology: Experimental setting

- In vitro experiment (27 participants, undergrads, Dpt. of Archives & Library Sciences)
- 13 given information needs related to environmental issues
  - The queries submitted for this purpose formed the terms' corpus
- Subject search only, Simple search
- Bibliographic database of the Evonymos Ecological Library (customized)
- Transaction Log Files (1 xml log file/ per user/ per session)
- Language used: Greek (highly inflectional)

Methodology: The Workflow



Methodology: Analysis specifications

- Manual designations to categories of failed queries, all examined one-by-one
- Use of Open Xerox tokenizer and morphological analyzer (PoS tags also included)
- Use of MS Word dictionary

♦ For more definitions and a full list of citations, please, refer to the full paper of this presentation hosted at the workshop's website

Results (i)

- 1,284 queries were submitted overall
- 36% were failed queries (i.e. 459)
- Further categorization of Failed Queries

Valid terms with zero hits	*Typing errors*	*Inseparable terms*	*Undefined terms*
75.8 %	19.6 %	2.4 %	2.2 %

- Further categorization of Typing errors

Substitution	*Transposition*	*Omission*	*Insertion*	*Division*
36.7 %	4.4 %	28.9 %	28.9 %	1.1 %

Results (ii)

- Typing error queries: 90
- Result after tokenization: 156 tokens
- Morpho-syntactic analysis of tokens (terms as submitted): 20/156 identified and analyzed

Poor performance in identifying the terms

Categorization of identified tokens (terms as submitted)				
Regular words	*Punctuation*	*Pronouns*	*Preposition*	*Other*
10	5	3	1	1

Results (iii)

- Error correction of tokens using MS Word dictionary
  - At this point we interfered with the results assigning the semantically correct word

No suggestion needed	30.1%	47
No suggestion made	12.8%	20
Irrelevant suggestion	3.2%	5
MS Word's 1st suggestion=correct	45.5%	71
MS Word's 2nd suggestion=correct	7.1%	11
MS Word's 3rd suggestion=correct	1.3%	2
Total	100.0%	156

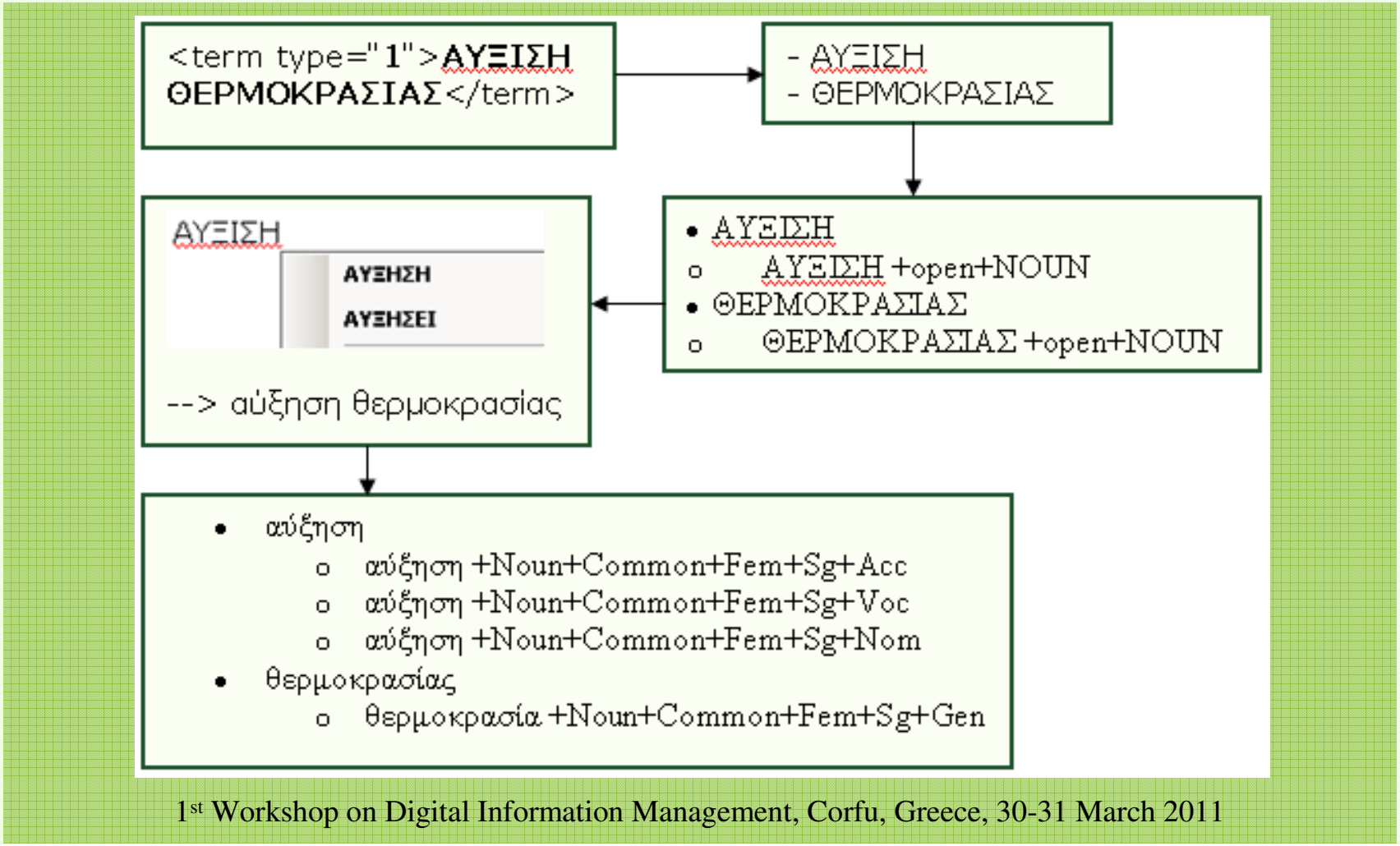
Results (iv)

- Morpho-syntactic analysis of corrected tokens: 139/156 identified and analyzed

Good performance in identifying the terms

Categorization of corrected tokens not analyzed		
Named entities	17.6%	3
Regular words	35.3%	6
Truncated words	29.4%	5
English words	11.8%	2
Punctuation	5.9%	1
Total	100.0%	17

An example


<term type="1">ΑΥΞΙΣΗ ΘΕΡΜΟΚΡΑΣΙΑΣ</term>
- ΑΥΞΙΣΗ
- ΘΕΡΜΟΚΡΑΣΙΑΣ
• ΑΥΞΙΣΗ
o ΑΥΞΙΣΗ +open+NOUN
• ΘΕΡΜΟΚΡΑΣΙΑΣ
o ΘΕΡΜΟΚΡΑΣΙΑΣ +open+NOUN
ΑΥΞΙΣΗ
ΑΥΞΗΣΗ
ΑΥΞΗΣΕΙ
--> αύξηση θερμοκρασίας
• αύξηση
o αύξηση +Noun+Common+Fem+Sg+Acc
o αύξηση +Noun+Common+Fem+Sg+Voc
o αύξηση +Noun+Common+Fem+Sg+Nom
• θερμοκρασίας
o θερμοκρασία +Noun+Common+Fem+Sg+Gen

Conclusions (i)

- 36% of the queries submitted returned zero hits
- 19.6% of failed queries were due to typing errors
- Typing error queries require more steps in the process towards morpho-syntactic analysis
- Tools for morpho-syntactic analysis of the Greek language need to be rich in tags in order to work adequately. This affects the complexity of the tool but it is essential since Greek is a highly inflectional language and cannot be analyzed easily.
- Transaction Log files offer a good starting point but need further analysis and support from other tools in order to be used for successful word-sense disambiguation

Conclusions (ii)

- MS Word dictionary
  - Was consulted in order to correct ~70% of the typing errors. The remaining were the part of submitted phrases with no typing error.
  - In 45.5% of the cases, first MS Word suggestion was correct.
  - Seems to favor at all instances "plural masculine" over "singular feminine", i.e. for "πυρινική", it first suggests "πυρηνικοί" and second "πυρηνική".
  - Does not recognize named entities

Conclusions (iii)

- Open Xerox
  - Works well for identifying and characterizing the segments
  - Is not enriched with many tags which leads to multiple suggestions. The goal in these cases is the less possible suggestions for each segment
  - Does not recognize capitalized words or words with no accent marks, meaning that the text submitted must be preprocessed to meet certain requirements
  - Does not recognize named entities

Future Work

- Deal with matters such as
  - Named entities recognition
  - Language identification
  - Word-sense disambiguation

  in order to achieve higher rates of morpho-syntactic analysis and apply automated mechanisms in this process
- Match the analyzed corpus of queries to Knowledge Organization Systems to assist query expansion

Thank you for your attention!

Any questions

Contact: Anna Mastora
Laboratory on Digital Libraries & Electronic Publishing,
Department of Archives & Library Sciences, Ionian University,
72 Ioannou Theotoki str., 49100, Corfu, Greece
mastora@ionio.gr