



Query Expansion and Context: thoughts on Language, Meaning and Knowledge Organisation

Anna Mastora & Sarantos Kapidakis

Laboratory on Digital Libraries and Electronic Publishing

Department of Archives and Library Science

Ionian University

2nd Workshop on Digital Information Management

25-26 April 2012, Corfu, Greece

Presentation Outline

- ▶ Introduction
- ▶ Research hypothesis
- ▶ Language & Problems in Information Retrieval
- ▶ Query Expansion
- ▶ Context
- ▶ Knowledge Organisation Systems (KOS)
- ▶ Considerations

Introduction

- ▶ Information Retrieval (IR) is about retrieving relevant results as response to expressed information needs
- ▶ The expression of information needs uses natural language representations
 - Language is clearly ambiguous (Szostak, 2010)
 - Not all linguistic representations are distinguishably informative of the user's intent

Representational indeterminacy

- ▶ Description of terms
 - Deals with each term individually
 - e.g. definitions in a dictionary

- ▶ Discrimination of terms
 - Deals with the relations between terms
 - e.g. hierarchical relationships within a thesaurus

(Blair, 2006)

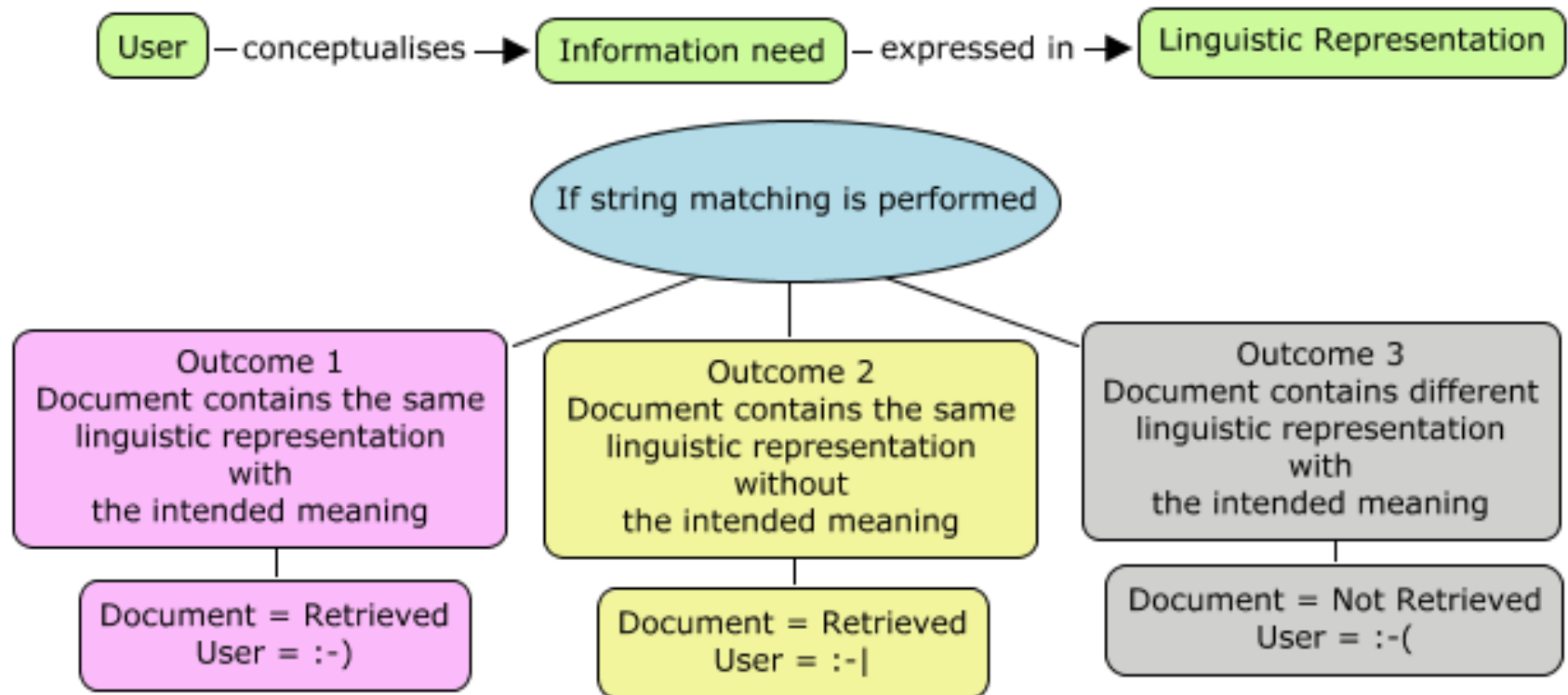
Research hypothesis

- ▶ By studying the actual use of language in every-day activities is how we can clarify meaning; deriving information about the context within which words are used will help us clarify better the identified indeterminacies

To strengthen the hypothesis

- ▶ *[...]for it's not an underlying logic that clarifies what we mean, it's the context, activities and practices in which we use language that provide the fundamental clarification of meaning we are looking for* (Blair, 2006)
- ▶ *Only in the stream of thought and life do words have meaning.* (L. Wittgenstein, "Zettel", §173)
- ▶ *Meaning depends on consistent usage but requires more than that; it also requires that speakers be able to check that someone's usage is consistent* (Jaworski, 2011)

Problems in IR due to uses of Language: Query – Document terms mismatch problem



Problems in IR due to uses of Language


Language is a labyrinth of paths. You approach from one side and know your way about; you approach the same place from another side and no longer know your way about. (§203)

L. Wittgenstein
“Philosophical Investigations” (1967)

Query expansion

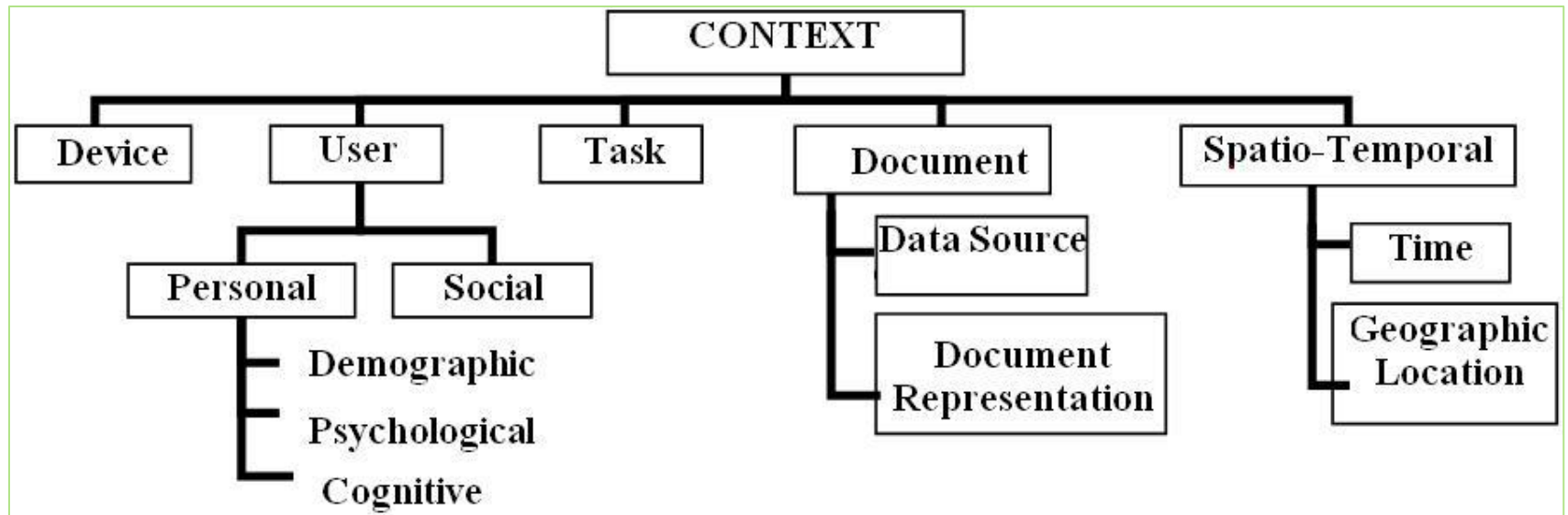
- ▶ In order for Information Systems to deal with the problems caused by the diversity of linguistic representations, Query Expansion is implemented.



- Supplementing the original query with additional – meaningful– words or phrases (manually, automatically, semi-automatically)
 - What *meaningful* means?
 - More information about the “context”
 - What is the *context*? 

Context: what is it?

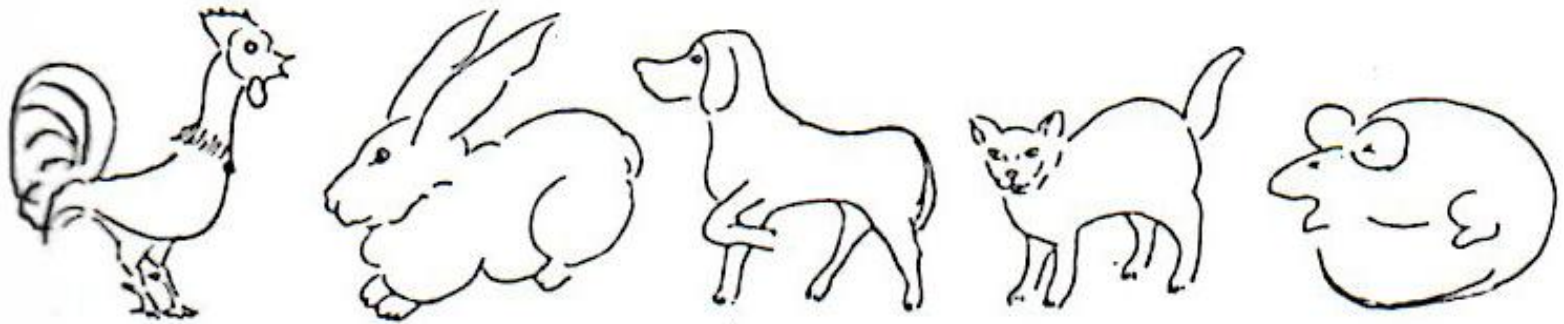
More information about the involved parties within the information retrieval process



The Multi-faceted concept of Context (Bhatia & Kumar, 2010)

Context: why is it important?

What do you see below?

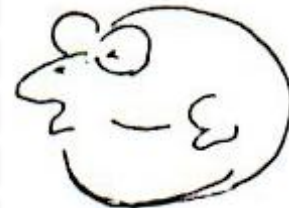
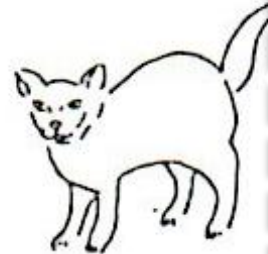
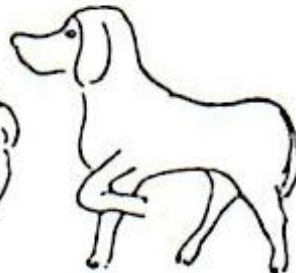
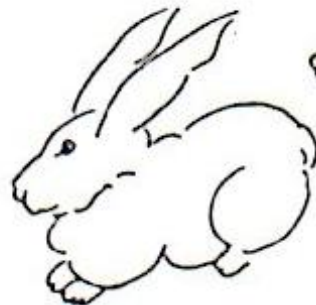


Context: why is it important?

What do you see below?



Context: it is important



(Porpodas, 1991)

Context: how do we get information about it?

- asking the users explicitly
- statistical processing of log files (e.g. Query chains)
- qualitative evaluation of log files, data or systems
- pre-processing of document corpora
- machine learning techniques (un-, semi-, supervised)
- implementation of user behaviour models
- personalisation/ profiling of users
- knowledge organisation structures
- ...

The indexer (i)

- ▶ Meant to be some kind of a problem solver; a mediator between the inquirer and the content which is stored in the documents
- *Placing a concept within a hierarchical definition establishes what sort of thing this is and what sort of thing it is not, and often sorts the subsidiary elements of which it may be comprised* (Szostak, 2010)

The indexer (ii)

The expert does not hold a “more complete” definition of the words than we do. He simply knows more about certain words than we do, and by providing this additional information about them, it may be useful for identifying each word in different circumstances.

(Blair, 2006)

Knowledge Organisation Systems (i)

- ▶ Schemes for organising information & promoting knowledge management (Hodge, 2000)
 - Term lists (authority files, glossaries, dictionaries, gazetteers)
 - Classification & categories (classification scheme, taxonomy, subject headings)
 - Relationship lists (thesaurus, semantic network, ontology)
- ▶ Constitute ways for formalising knowledge and, consequently, communication through linguistic expressions or any other kind of definitional or descriptive sign, like visual.
- ▶ Used in Query Expansion to derive context

Knowledge Organisation Systems (ii)

- ▶ *The epistemological basis of any theory of Knowledge Organization is an accepted postulate. In other words, how knowledge is organized and represented depends largely on the understanding of how knowledge is organized and represented.*

(Alexiev & Marksby, 2010)

Knowledge Organisation Systems (iii)

- ▶ Since language is involved in their creation and development, KOSs bear themselves the inherent characteristics of the use of language
 - Ambiguity [PoS: “looks”, Semantic: “bank”, Syntactic: “He hit the girl with the hat”]
 - Homonymy [Homophones: “too /two”, Homographs: “tire”]
 - Polysemy [“mouth” (on the face OR the opening of a cave)]
 - Synonymy [“sick – ill”]

Knowledge Organisation Systems (iv)

- ▶ Even if KOSs try to capture and deliver the absolute meaning [or a more targeted one] of what they describe, they still are considered “*collection independent knowledge structures*” (Efthimiadis, 1996)
 - There still are missing parts for the communication of the intended meaning

Ambiguity differs only by degree between universal and domain-specific classifications, though that difference of degree is likely quite significant

(Szostak, 2010)

Considerations

- ▶ Take advantage of *user models*
- ▶ Take advantage of *user evaluation*



They deliver information about the real use of language

The degree of ambiguity lessens within groups that regularly interact (though it does not disappear)

(Szostak, 2010)

Thank you!

Contact:
Anna Mastora
mastora [at] ionio [dot] 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.

