INNOVATION, LANGUAGE, and the WEB

Claudia Marzi

Institute for Computational Linguistics "Antonio Zampolli" – CNR
University of Pavia – Dept. of Theoretical and Applied Linguistics

LANGUAGE and INNOVATION are inseparable. Language conveys ideas which are essential in innovation. Every linguistic choice is meaningful, and it involves the parallel construction of form and meaning.
→ Language is a dynamic knowledge construction process.

The importance of efficiently deploying knowledge for a complete and successful exchange is easily understandable: through a better understanding of information new ideas can be captured and exploited. Language has effects at all stages of knowledge transfer. Knowledge ambiguity may also depend on language ambiguity.

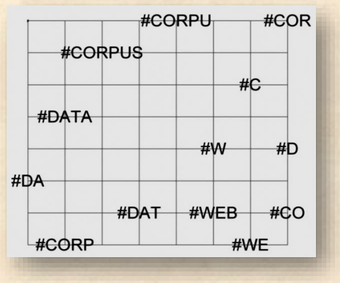
Our goal is to focus on how words and language structures become vehicle for knowledge generation, in particular for innovation transfer.

→ Linguistic innovations arise in the context of existing rules which they modify.

- LANGUAGE AMBIGUITY → Semantically ambiguous and polysemous words can be disambiguated by defining the context; polysemous words, in particular, shape their meaning as a function of their context of use.
- Methods for computing relational similarities and disambiguating polysemous words, based on large text corpora, can make rough sense distinctions, though remaining far from reaching the sophistication of human judgement.

However, register distinctions are not defined in linguistic terms, but rest on context, domain, and purpose; and contextual knowledge allows to support KNOWLEDGE PROCESSES and to better access them.

- The meaning of words can change over time and discourse; words can take on new senses when used in novel context; words with emergent novel senses often reflect an extension of use from one domain to another.
- In NLP collocational information derived from corpora is useful in the perspective of text analysis → in word sense disambiguation collocations are used to discriminate between sense of polysemous words. Frequently occurring collocates give the idea of semantic preference. Corpus data can be considered as very useful for revealing typically lexico-grammatical patterns and functional aspects of language.



The WEB as a linguistic corpus → to investigate how words are used to describe innovation, and how innovation topics can influence word usage and collocational behaviour. As a source of machine readable texts the Web offers a huge repository of documents written in a multitude of languages
→ *Problem* → far from standard, they contain different types/genres, and constantly change over time.

The proposed study is based on NLP technologies → a computational approach to words sense disambiguation is identified by focusing on similarity in context.

METHODOLOGY

Genre-and domain-oriented texts are analysed with the support of SKETCH ENGINE (http://www.sketchengine.co.uk)
→ Corpus query tool, based on a distributed infrastructure, that generates word sketches and a thesauri.
→ Polysemous test words have been selected with reference to innovative domains and their collocations are analysed.

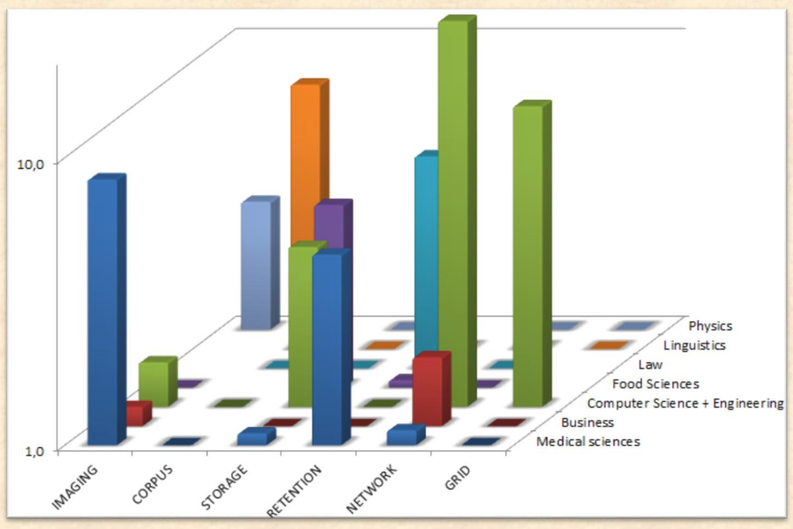
Test words present a potentially high degree of semantic ambiguity/polysemy and different context collocations:
IMAGING, RETENTION, STORAGE, CORPUS, NETWORK, GRID.

Texts materials: selected by using search web engine queries (www.google.com by using seed words), domain-specific databases and type coherent multidisciplinary large corpora (www.opengrey.eu, www.ncbi.nlm.nih.gov/pubmed by selecting the domain). Collocations and concordances are compared with balanced corpora (e.g. the British National Corpus, British Academic Written English).

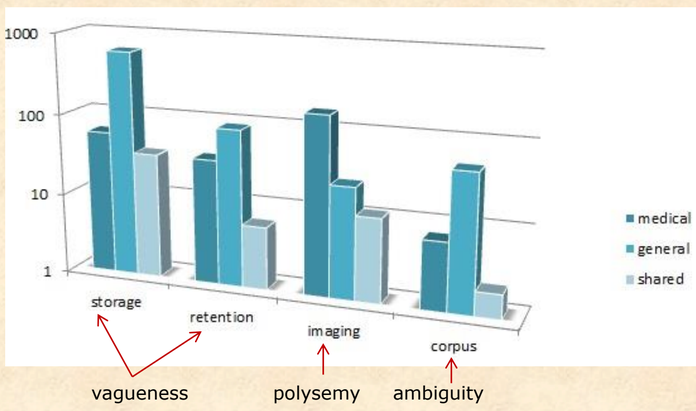
RESULTS

By comparing collocates and keywords in different contexts of use, we investigate the ambiguity vs. polysemy gradient, showing how dynamically word meanings are adjusted to novel usages

All six terms exhibit distinct senses when used in different contexts:



- CORPUS can refer to brain areas (medical domain) as well as large collection of items (general sense)
- IMAGING means visual representation, but in the medical domain refers to specific diagnostic technology
- NETWORK and GRID represent somewhat extreme cases of such domain-sensitive specialisation, to the point that they appear to be overwhelmingly used in their specialised senses only
- RETENTION and STORAGE byselecting both material and immaterial items appear to oscillate between their proper and extended senses interchangeably, witnessing a paradigmatic case of systematic, context-sensitive polysemy.



The lexical representation of INNOVATIVE KNOWLEDGE requires a dynamic shift from context-driven vagueness (semantic polymorphism) to domain-driven specialisation (terminological usage)