Consiglio Nazionale delle Ricerche

INNOVATION, LANGUAGE, and the WEB

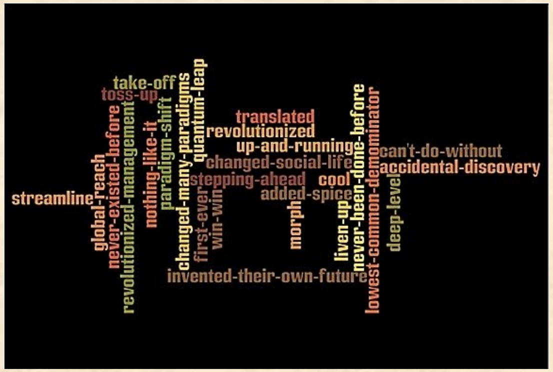
Claudia Marzi

Institute for Computational Linguistics, “Antonio Zampolli” – Italian National Research Council

University of Pavia – Dept. of Theoretical and Applied Linguistics

DYNAMICS OF LANGUAGE

INTRODUCTION



Language conveys ideas which are essential in corporate innovation; innovation would be nearly impossible if we did not have language.

Language makes us shape and refer to things, events, and concepts.

Communication takes place when there is a real information exchange process:

- every linguistic choice is necessarily meaningful
- language is in this sense a dynamic communicative and interactive process

RESEARCH GOAL

The communicative ability connects people into information-sharing network and information allows people to expand their knowledge.
In this context, our goal is to:

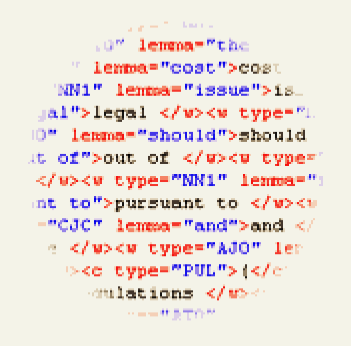
➡ focus on how words and language structures become vehicle for knowledge generation, and in particular for innovation transfer.

➡ focus on lexical creativity, and its dynamic interplay with innovative contexts

WEB AS A CORPUS

The WEB:

- has become a primary meeting place for information exchange;
- has evolved into a primary resource for lexicographers and linguists;
- Is a source of machine readable texts for corpus linguists and researchers in the fields of Natural Language Processing (NLP), Information Retrieval and Text Mining



```
O" lemma="the
" lemma="cost">cost
NN1" lemma="issue">is
al">legal </w><w type="
O" lemma="should">should
t of">out of </w><w type="
</w><w type="NN1" lemma="
nt to">pursuant to </w><w
="CJC" lemma="and">and </
e </w><w type="AJ0" le
><c type="PUL">(</c
ulations </w><
="AT0"
```

WEB searching as default mode for information retrieval, though the main sources of digital information are unstructured or semi-structured text materials

CORPUS LINGUISTICS

The meanings of words can change over time/discourse, and words can take on new senses

Large corpora – large and growing repositories - to investigate:

- how words are used to describe innovation
- how innovation-driven topics can influence word usage and collocational behaviour

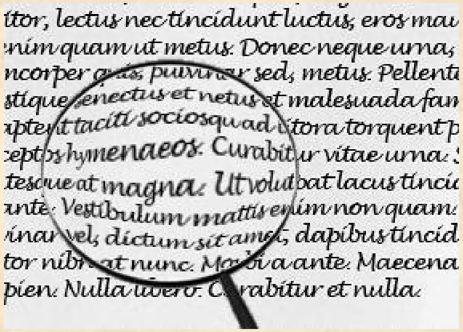
Corpus-based studies → lay emphasis on how word usage and word formation processes conveys differences in context, domain and purpose

RESEARCH RATIONALE

- The meanings of words can change over time, and words can take on new senses
- Words with emergent novel senses reflect an extension of use from one domain to another
- Semantically ambiguous words (polysemy) can be disambiguated by defining the context – what people do in spoken language (pragmatic competence)
- Analysis on language productivity offer the opportunity to investigate how words become vehicle for innovative knowledge generation and transfer

METHODOLOGY

Genre-oriented and stylistically heterogeneous texts are taken:



→ Bibliographic (domain-specific) database
→ Type coherent multidisciplinary database

Search web engine and large corpus query tools are used:
→ Generic search engine
→ Lexicography query system

Domain-specific vs. generic-purpose texts offer materials for terminology exploration

EXPERIMENTAL EVIDENCE

By comparing reference corpora – both specific and generic – with the web, examples of polysemy are given:

- ✓ *Imaging* = brain imaging techniques vs. visual representation → ambiguous
- ✓ *Neuro* = referred to diagnosis, medicine and surgery. → coherent
- → *Neuroimaging* → coherent

→ ambiguous

- ✓ *Storage* = memory capacity *vs.* containing units
- ✓ *Memory* = process of information encoding and retrieving → coherent

What about a polysemous term like *retention*? What domain does it belong to? What is the meaning and what concepts does it convey?

DYNAMICS OF LANGUAGE

DISCUSSION and CONCLUSION

- Lexical co-occurrences and collocations can be of considerable help in retrieving text materials which are relevant to a specific domain of interest; however they are often helpless in telling genres purposes or intended readerships

- Bibliographic domain-specific databases are developed to offer materials to a specialised readership, thus providing highly-selected, well-targeted documents

- We investigate the hypothesis that specific word-formation processes and neologisms correlate significantly with text genres and intended readership, thus providing a convenient way to track down well-targeted, highly technical repositories of openly available text materials