Consiglio Nazionale delle Ricerche

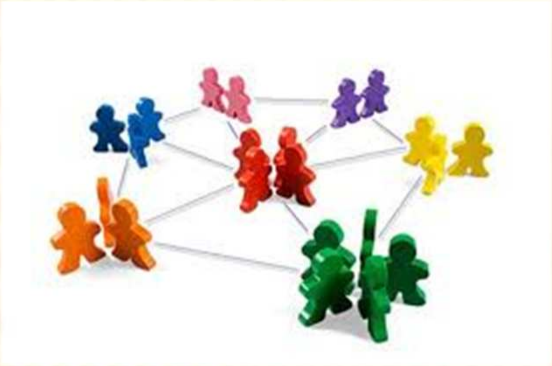
KNOWLEDGE COMMUNITIES IN GREY

Claudia Marzi

Institute for Computational Linguistics (ILC)
National Research Council (CNR) - Italy

DYNAMICS OF LANGUAGE

INTRODUCTION



The dynamic nature of modern human social interactions, and the increasing capability of wireless and mobile devices for creating and sharing contents, open up the opportunity for a wide dissemination of information through complex knowledge sharing systems.

The web offers a steadily increasing availability of ubiquitous accessible information.

In this context, social networks can enhance fore-front ideas and highly innovative contents;

⇒ they offer an enormous potential to transform research, and research results, into a knowledge co-creation process.

RESEARCH QUESTIONS

- To what extent can Social Networks provide a real opportunity for sharing knowledge and generating and disseminating novel information?

- Can they really be supportive of a steady flow of technical and scholar writing, or do they only provide a general communication channel for ephemeral communication exchanges?

- Is there a specific added value in the way Social Networking can foster people's interest in sharing and building information?

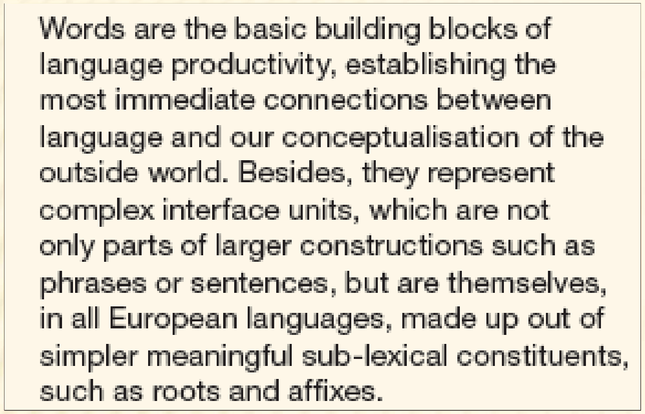
- Is interactive, informal and ubiquitous information exchange developing a new social framework for the creation of public-domain knowledge?

RESEARCH RATIONALE

We suggest that all these questions can be addressed by applying advanced NLP tools for automated content extraction to the analysis of web-based text collections, sampled from both general-purpose and specialized examples of social networks.

- The Information Extraction literature provides different modes and tools for knowledge acquisition and representation: from highly structured, standardized and objective knowledge information systems based on ontological hierarchies and relations to more dynamic, subjective tools for volatile knowledge representation such as word clouds and concept maps.
- Technologies in NL understanding offers an objective measure of the information density of a text document or document collection and ways to map out the distribution/development of information. This makes it possible to compare the information structure across texts and get a sense of their level of content sharing and knowledge coherence.

METHODOLOGICAL APPROACH

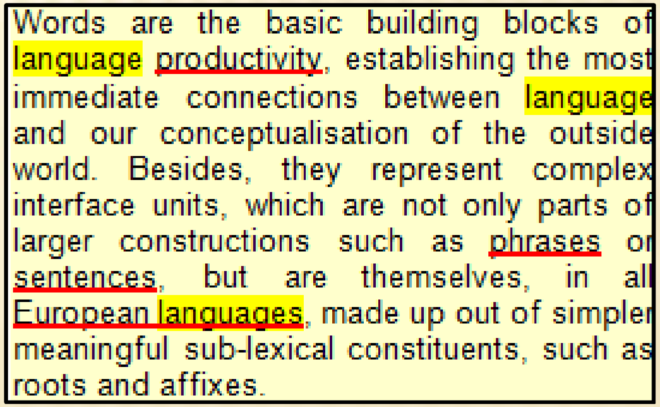


Words are the basic building blocks of language productivity, establishing the most immediate connections between language and our conceptualisation of the outside world. Besides, they represent complex interface units, which are not only parts of larger constructions such as phrases or sentences, but are themselves, in all European languages, made up out of simpler meaningful sub-lexical constituents, such as roots and affixes.

Natural Language Processing tools can augment text documents with layers of mark-up data, making the hidden linguistic structure of the document overtly represented and accessible

The input text is segmented down into words and multi-word structures, mutually linked through syntactic relations

Salient terms are identified in context, to provide access keys to the basic contents of the document



Words are the basic building blocks of language productivity, establishing the most immediate connections between language and our conceptualisation of the outside world. Besides, they represent complex interface units, which are not only parts of larger constructions such as phrases or sentences, but are themselves, in all European languages, made up out of simpler meaningful sub-lexical constituents, such as roots and affixes.

AUTOMATED CONTENT ANALYSIS

Linguistically annotated documents provide a jumping-off point for the acquisition of more and more abstract representations of the document content:

- words are structured into terms,
- terms are grouped into conceptual classes,
- concepts are linked together through vertical (taxonomical) and horizontal (ontological) relations.

This kind of linguistic information represents the basis of a computational platform for automated content sharing, access and dissemination.

AUTOMATED CONTENT ANALYSIS (II)

Moreover, through NLP technologies another orthogonal level of linguistic information can usefully be represented:

- the content accessibility – level of readability of a text on the basis on its processing difficulty

This type of analysis allows us to compare the information structure of different text collections and get a precise sense of their level of informativeness. In particular:

- lexical richness has to do with the lexicon of a text
- lexical density gives a measure of the rate at which the content is updated

IN DETAIL

- We can identify the most salient terms in a document and the degree of subject-specificity by comparing the frequency distribution with the frequency distribution of same terms in a balanced corpus.

- The syntactic complexity can be calculated on the basis of:
  - ✓ the average length of clauses
  - ✓ the ways words are arranged in context
  - ✓ the length of dependency chains
  - ✓ the word distance between head and dependant

- Text excerpts are sampled from:

  - general-purpose social networks (based on friendship relations and social proximity)
  - specialized subject-based communities (based on content sharing and supporting relationships)

EXPERIMENTAL EVIDENCE

In the English experiment, we conducted a grammatical and content-word evaluation in 3 different text collections:

- ✓ a sample of messages posted in general-purpose social networks (e.g. Facebook),
- ✓ a sample of message exchanges within subject-based web communities (e.g. LinkedIn),
- ✓ as a base-line, a sample of Grey Literature writings (GL 12 Conference Proceedings).

The distribution of terms was comparatively evaluated on the basis of the degree of domain-specificity of terms

	domain specific terms (single and multiple)
Social networks	1.75
Subject-based communities	14.75
GL Papers	15.75

Text analysis tools: http://www.ilc.cnr.it/dylanlab/

EXPERIMENTAL EVIDENCE (II)

A measure of the average syntactic complexity is given, resulting form:

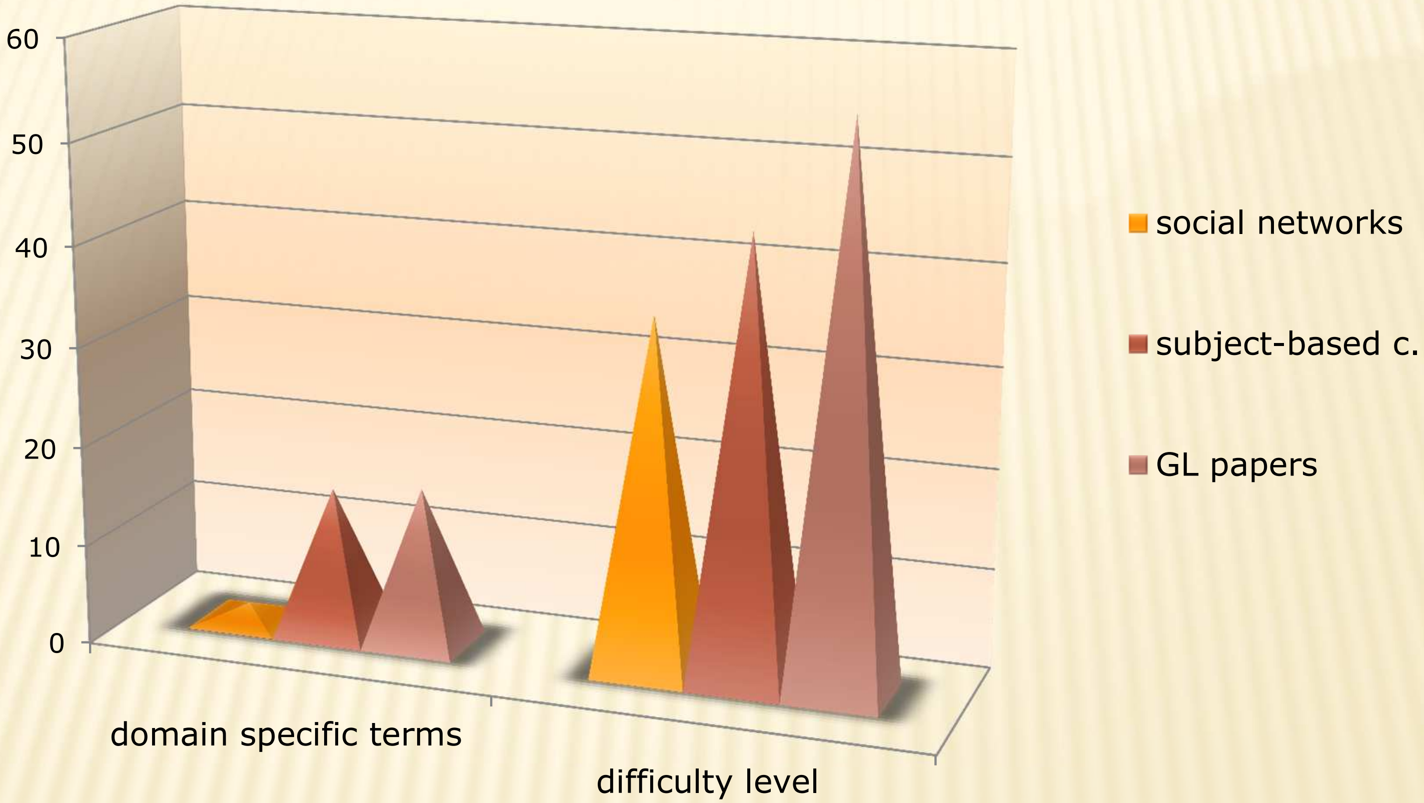
✓ lexical rarity of content-words,
✓distribution of part-of-speech tags
✓average length of chains of dependency links
✓average head-complement distance

Once more, the 3 text samples, ranked by increasing values of syntactic difficulty, reflect a gradient of content accessibility which appears to mirror the degree of communicative formality (from less formal to more formal) scored in our text types

	difficulty level
Social networks	35.30
Subject-based communities	43.85
GL Papers	55.20

Text analysis tools: http://www.ilc.cnr.it/dylanlab/

RESULTS



Text analysis tools: http://www.ilc.cnr.it/dylanlab/

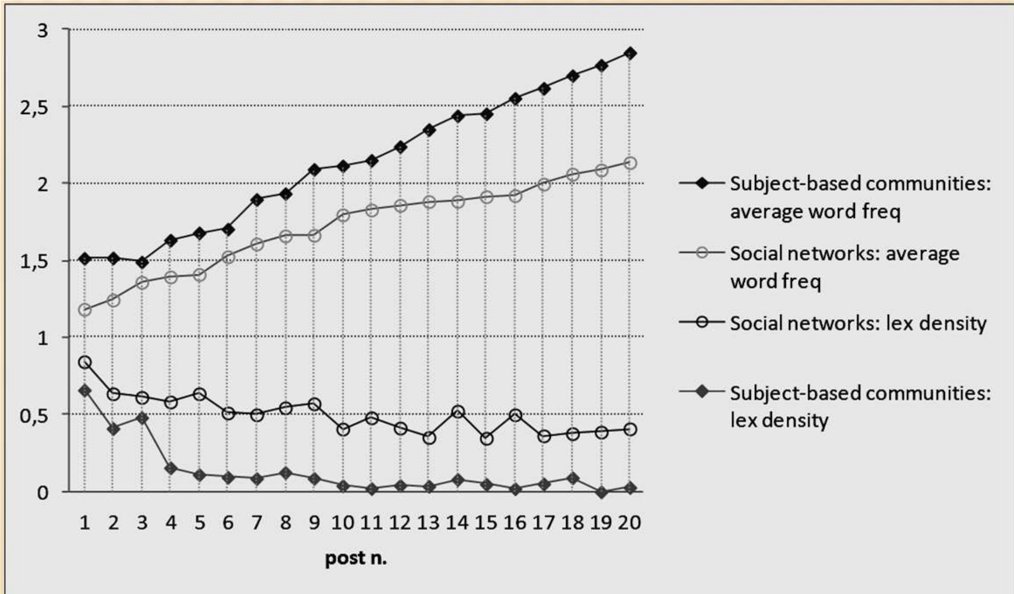
EXPERIMENTAL EVIDENCE

In the Italian experiment, we compared the overall levels of lexical coherence in post exchanges of various fb accounts' contacts based on friendship relations, and of a subject based blogs of CNR intranet.

Lexical coherence is automatically estimated by measuring the flow of new lexical items that are incrementally added in a post exchange referring to the same issue -calculated as the number of novel words introduced by each newly posted comment divided by the length of the comment.

Social networks show a slower growth of average word frequency;

Subject-based writing tends to be lexically more coherent



DISCUSSION

Social networks

The medium of Social Networks tends to make communication simpler, with:

- shorter sentences than in traditional texts
- high/medium-frequency words
- simpler syntax (one verb per sentence)
- high readability

with no guarantee of knowledge sharing

Subject-based communities

In case of stronger interaction between medium and content, we observe:

- domain-specific terms
- high lexical coherence
- more levels of syntactic embedding
- high/medium readability

CONCLUDING REMARKS (I)

NLP tools for content analysis and Information Extraction establish a direct relation between modes of knowledge creation/sharing and dynamic, incremental approaches to automated knowledge acquisition and representation:

- they allow us to assess the content of a text in terms of its level of readability, domain-specificity, lexical coherence and density of its conceptual maps;
- they can be used to measure not only the effectiveness of a text in conveying information but also the extent to which this information is structured in terms of shared knowledge.

CONCLUDING REMARKS (II)

- **Subject based communities**, focused on supporting relationships and content sharing, act at the same time as providers and users of all kind of GL materials in a highly distributed and collaborative scenario, and represent a conducive environment for knowledge transfer.

  They can represent – as **collaboration networks** – a key element in the advancement and dissemination of knowledge in scientific domains as well as in diverse aspects of everyday human life.

- General-purpose **social networks**, reflecting either friendship or superficial relationships, tend to generate ephemeral information and to create a more superficial and mosaic knowledge.

Social Networking is a medium with a strong potential, a house of cards powerful and fragile at the same time