

## KNOWLEDGE COMMUNITIES IN GREY

Claudia Marzi

*Institute for Computational Linguistics, "Antonio Zampollì"*

*National Research Council, Italy (CNR)*

Via G. Moruzzi 1, 56124 Pisa, Italia

claudia.marzi@ilc.cnr.it

### Abstract

*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.*

*As the shared knowledge components build cognitive ties, there is no real sharing of knowledge without a common understanding of it.*

*In this article, particular emphasis is laid on technologies in Natural Language understanding and knowledge management for providing structured, intelligent access to the continuously evolving content, generated on-line in a pervasive collaborative environment.*

*In detail, robust automated techniques for term extraction and knowledge acquisition are used to tap the information density and the global coherence of text excerpts sampled from both general-purpose and subject-specific social networks. We show empirically that the two sources may exhibit considerable differences in terms of content accessibility and informativeness.*

**Topics: Subject based Communities; Social Networking**

**Keywords: Grey Literature, Web Communities, Knowledge sharing, Concept Maps**

### 1. Introduction

The development of digital technologies and the continuous evolution of telecommunication networks are rapidly heading our society towards a culture of participation and to a more and more interactive communication. The dynamic nature of modern human social interactions, the adaptive networking protocols and data management systems are fostering pervasive information and communication environments.

The web represents an unlimited universe of information and data, and offers the steadily increasing availability of ubiquitous accessible information.

As accessibility improves, however, the huge amount of data and information available on the web need to be identified, classified, analyzed, filtered, so as to enhance the generation and assimilation of new knowledge.

Large volumes of information, even structured information, have to be managed, and generation and assimilation of knowledge have to be facilitated. Knowledge needs to be represented, standardized and distilled from multiple sources.

In this context, Social Networks can enhance fore-front ideas and highly innovative contents; they offer the enormous potential to transform research, and research results, into a knowledge co-creation process.

## **2. Methodology**

### 2.1 Research questions

Given this scenario, the following questions arise naturally.

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

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?

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

### 2.2 Research rationale

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 Natural Language understanding offer 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.

This approach will highlight current automated tools for concept acquisition and ontology learning that are conducive to an incremental access and management of content, to establish a fruitful bridge between modes of knowledge sharing/creation and dynamic, incremental approaches to automated knowledge acquisition and representation.

### 2.3 Methodological approach

Natural Language Processing (NLP) 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. Moreover, salient terms are identified in context, to provide access keys to the basic contents of the document. In classical NLP architectures, this is carried out in a step-wise fashion, with layers of annotation being cascaded in a feeding relation, from *tokenized* texts to trees of dependency relations. Typical parsing steps are: i) *tokenization*, ii) *morphological parsing* and iii) *dependency trees*.



*Tokenization* amounts to assigning a string of characters the status of single token, where a token is the most basic parsing unit, approximately corresponding to a linguistic word, but also including non-lexical units such as dates, addresses, proper names, acronyms, measuring expressions, etc. For tokens to be identified as independent words and assigned their corresponding part-of-speech tag (or grammatical category), their set of morpho-syntactic features (e.g. number, gender, tense, etc.) and their lemma (or lexical exponent), they have to undergo a level of *morphological parsing*. In its simplest instantiation, morphological parsing requires the existence of large repositories of word forms, where each form is glossed with a set of morpho-lexical features. However, in

languages with rich morphologies, a closed-list approach to morphological parsing is subject to serious risks of failure, as shown by the German example in (1) (borrowed from Anderson and Lightfoot, 2002):

- (1) *Lebensversicherungsgesellschaftsangestellter*  
(*life insurance company employee*)

In fact, no German lexical repository can be expected to be large enough to contain all possible compounds of this kind. A principled solution is to split the compound into its simpler constituent words (*Leben + Versicherung + Gesellschaft + Angestellter*), for the latter to be looked up in a lexical database as individual entries.

Once word tokens are identified and categorised for contextually-appropriate part-of-speech and lexical exponence, they are grouped into larger constituents defining their *syntactic dependency relations*. A dependency relation is a binary relation linking two tokens in context, usually represented as a pointed arc going from the *dependant token* (a complement or a modifier) to its syntactic head (usually a complemented or modified verb or noun). Dependency relations can also be defined between the constituents of complex NN compounds as shown by the following examples:

- (2) a. *life insurance company employee*  
  
b. *china tea cup*  


The pointed arcs above tell us that *life insurance* is a type of *insurance*, *life insurance company* is a type of *company* and *life insurance company employee* is in fact an *employee*. Incidentally, it should be appreciated that not all dependency chains in NN compounds must look like those in *life insurance company employee*, as shown by the dependency structure of *china tea cup* above, where both *china* and *tea* entertain a modifying relationship with *cup*.

Linguistically annotated documents provide a jumping-off point for the acquisition of a more and more abstract representation of the document content, in line with the so-called “layer cake” approach to ontology learning (Buitelaar, Cimiano and Magnini 2005), whereby:

- ✓ words are structured into terms (e.g. *life insurance company* is a complex term),
- ✓ terms are grouped into conceptual classes (e.g. *insurance company* is a co-hyponym of *telecommunication company*),
- ✓ concepts are linked together through vertical (taxonomical, e.g. *life insurance company* is a hyponym of *company*) and horizontal (ontological, e.g. people are typically employed in a *company*) relations.

Such a wealth of information provides the basis to a computational platform for automated document content sharing, access and dissemination, that allows document contents to be queried by concepts and concept relations rather than by fixed text patterns or key-words (Lenci et al., 2008). For example we can search a text for information about the number of employees of a given insurance company or its overall yearly income and the like. With no linguistic information such as in (2.a-b) above, a text can be navigated only through fixed word patterns. Linguistic annotation offers a more abstract level of information which can selectively be searched for intelligent information access.

Another, orthogonal level of linguistic information that can usefully be represented through NLP technologies is the *content accessibility* of a document, defined as the level of readability of a text document calculated on the basis of its processing difficulty. Processing difficulty is a multi-factorial concept, which can be decomposed into several, fairly independent factors, such as lexical richness, lexical density and (morpho-)syntactic complexity. Each factor can be assessed independently through measurable parameters. This type of analysis allows us to get an objective measure of the information density of a text document or document collection and to map out its distribution/development through the document(s). This makes it possible to compare the information structure of different text collections and get a precise sense of their level of informativeness, content sharing and knowledge coherence.

In particular, *lexical richness* has to do with the lexicon of a text document, defined as the set of word types attested in the document. Trivially, a richer lexicon has a higher set cardinality than a

poorer one. More subtly, lexical richness also involves word frequency distributions. Rare words are in fact taken to be more difficult to process than more common words and often denote the most salient pieces of content information of a document together with its level of subject-specificity. Accordingly, by inspecting the tails in the Zipfian distribution of different document lexicons we can get a flavour of the different degrees of lexical richness of the corresponding documents (or document collections). *Lexical density*, on the other hand, gives a measure of the rate at which the content of a collection is updated through the introduction of novel concepts and is defined as the number of new words that a text excerpt introduces in a document collection, divided by the length of the excerpt.

Salient domain-specific concepts and relations are most often conveyed in text through statistically significant terms. Relevant terminological units can be tracked down automatically by projecting abstract morpho-syntactic patterns such as “NP PP” (*i.e.* “find a syntactic structure made up out of a Noun Phrase immediately followed by a Prepositional Phrase) onto linguistically annotated texts. Text strings fitting into the targeted morpho-syntactic pattern are then filtered out through a further step of statistical post-processing, to assess their potential for termhood. Since Smadja’s (1993) seminal work, statistical methods offer reliable means of acquiring domain specific expectations concerning the joint distribution of words in sufficiently large training corpora. Association measures such as Pointwise Mutual Information (Church and Hanks, 1989) have become standard utilities to measure the degree of collocational association of word pairs in context, by exploiting the intuition that words belonging to the same bracketed pair will co-occur in corpora significantly more often than what would be expected under a model of chance co-occurrence (based on the frequency of the individual words). Based on this intuition, we can further identify the most salient terms attested in a document collection and their degree of subject-specificity, by comparing their frequency distribution in the target collection with the distribution of the same terms in a balanced, general-purpose corpus.

Finally, a score of (morpho-)syntactic complexity can be calculated on the basis of i) the average length of text clauses (the longer a clause, the more difficult to parse), ii) the way words are

arranged in context (an unusual/marked order of words is more difficult than a standard word order), iii) the per-sentence length of the attested dependency chains (shorter chains are easier to be parsed and understood), and iv) the per-word distance between a head and its dependant (the longer the distance the more difficult the relationship) (Dell'Orletta et al., 2011).

This multi-factorial information allowed us to compare the information density of two different text collections: samples of text excerpts from general-purpose social networks, based on friendship relations and on social proximity, and samples of texts produced within the frame of specialized subject-based communities, based on content sharing and supporting relationships.

### **3. Experimental evidence**

#### 3.1 Results

Two distinct experiments were carried out on English and Italian texts. For all experiments, we used the web-based battery of NLP and knowledge-management tools made available on-line by the *Dylan Lab*<sup>1</sup> (Dynamics of Language Laboratory, Institute for Computational Linguistics - Italian Research Council).

In the English experiment, we conducted a cross-evaluation assessment of both grammatical and content-word parameters in three different text collections: i) a sample of messages posted in general-purpose social networks (e.g. Facebook), ii) a sample of message exchanges within subject-based web communities (e.g. LinkedIn), iii) as a base-line, a sample of Grey Literature writings (e.g. *GL 12 Conference Proceedings*). The distribution of terms in the three samples was comparatively evaluated on the basis of the degree of domain-specificity of terms. This is shown in table 1, which gives the average number of terms that were automatically found to be domain-specific by comparison with a general purpose corpus of English language newspapers (*The Wall Street Journal* section of the Penn Treebank). Characteristically, texts exchanged by subject-centred communities contain an average number of domain-specific terms comparable to the

---

<sup>1</sup> Tools available at [http://www.ilc.cnr.it/dylanlab/index.php?page=software&hl=it\\_IT](http://www.ilc.cnr.it/dylanlab/index.php?page=software&hl=it_IT)  
[http://www.ilc.cnr.it/dylanlab/index.php?page=software&hl=en\\_US](http://www.ilc.cnr.it/dylanlab/index.php?page=software&hl=en_US)

number of terms conveyed on average by GL papers, and considerably higher than the average number of terms occurring in general social network comments.

**Table 1**

	<i>domain specific terms (single and multiple)</i>
Social networks <sup>2</sup>	1.75
Subject-based communities <sup>3</sup>	14.75
GL Papers	15.75

On the other hand, table 2 gives a measure of the average syntactic complexity of our samples, resulting from the weighted integration of several quantitative parameters: lexical rarity of content-words, distribution of part-of-speech tags (an objective measure of morpho-syntactic complexity), average length of chains of dependency links stemming from a single syntactic head, and average head-complement distance measured by the number of intervening words (Dell’Orletta et al. 2011):

**Table 2**

	<i>difficulty level</i>
Social networks	35.30
Subject-based communities	43.85
GL Papers	55.20

Once more, the three 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 (Figure 1).

<sup>2</sup> Facebook excerpts are messages posted during a time slot of 20 days on various accounts, varying in age, social class and professional level. The figure in the table is an average over excerpt samples whose standard deviation is 1.24

<sup>3</sup> LinkedIn excerpts are messages taken from the GreyNet Group, during a tile slot of 20 working days.



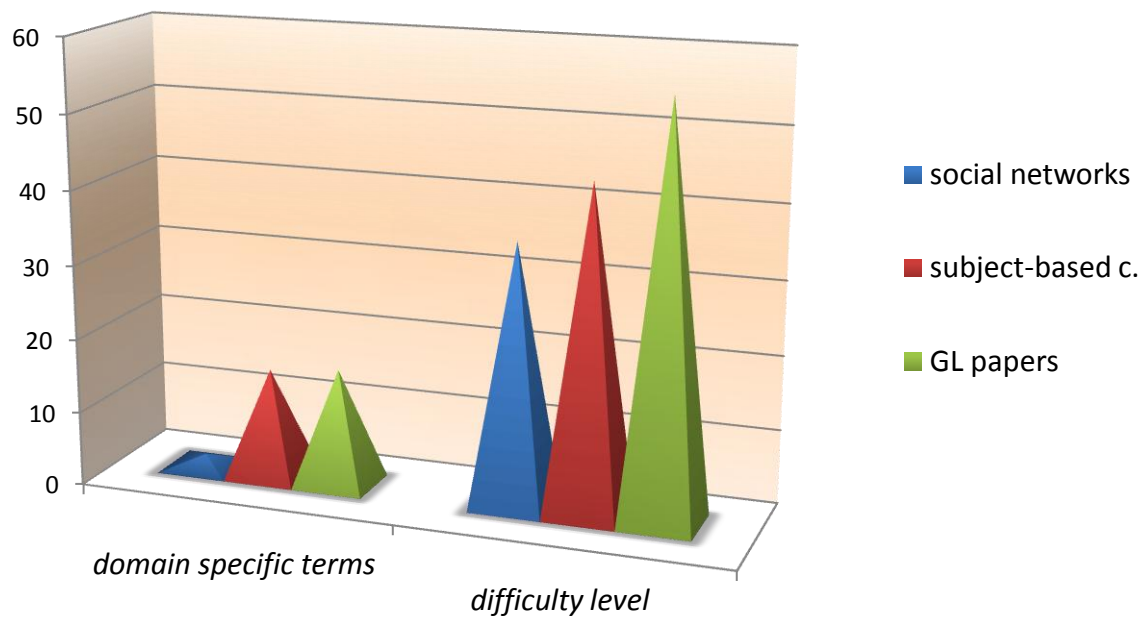


Figure 1

In the Italian experiment, we compared the overall levels of lexical coherence in two samples of Italian post exchanges: i) through a general-purpose social network (various Facebook accounts' contacts all based on friendship relations); ii) through subject-based technical blogs of a research institution (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. This is calculated as the number of novel words introduced by each newly posted comment divided by the length of the comment. Results are summarized in the graph below (Figure 2), providing the overall trend in the average word frequency and lexical density for the two text samples (over an exchange of maximum 20 posted comments).

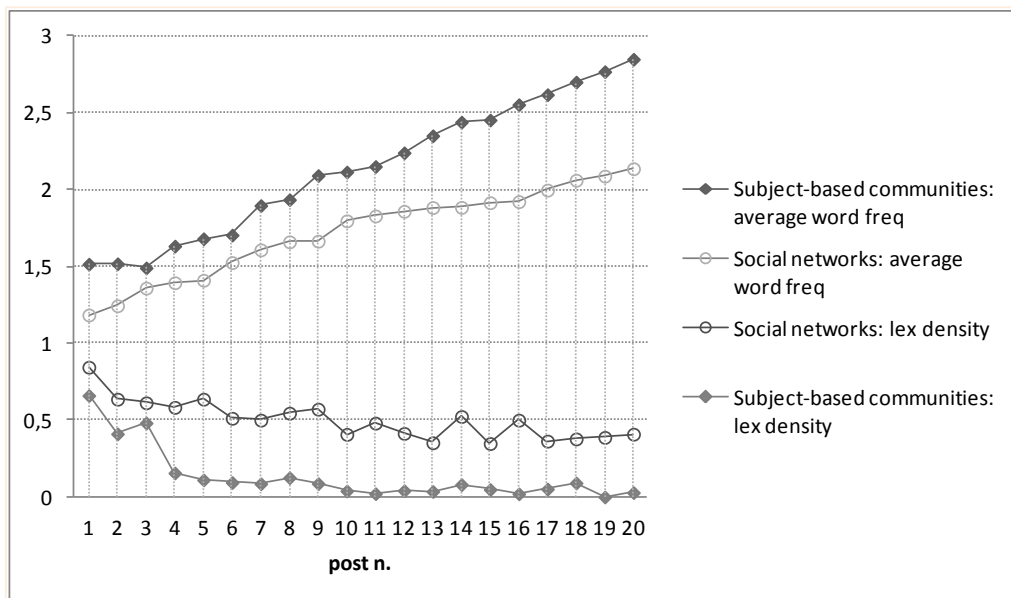


Figure 2

The two trends are remarkably different. Social networks show a slower rate of average word frequency, witnessing a higher variety of lexical choices (the same words are repeated less often). On the other hand, subject based technical writing tends to be lexically more coherent, with a systematic trend towards repetition of the same words. That these words are mostly technical is shown by the average number of domain-specific terms and their high level of difficulty/rarity as shown in table 3:

Table 3

	<i>Single terms</i>	<i>Multiple terms</i>	<i>Average domain specific terms</i>	<i>lexicon difficult level</i>	<i>Readability level</i>
Subject based CNR intra-blog	20	5	12.5	50.2	84.7
Social networks	1	3	2	12.5	73.7

### 3.2 Discussion

The two experiments were intended to test the empirical hypothesis that only subject-based collections offer a coherent flow of shared and structured knowledge, general purpose social networks being more erratic and ephemeral in the choice of discussion topics and domains. This hypothesis is basically confirmed by our results. The medium of Social Networks tends to make

communication simpler, with shorter sentences than in traditional texts, medium/highly frequency distributed words, simpler syntax (one verb per sentence), and a high readability score. This is true of all texts that were sampled from writings exchanged through social networks, irrespective of their topic.

However, a simpler communication does not necessarily guarantee coherence of information flow and steady knowledge sharing. Only in those cases where there is a strong interaction between medium and content (as in subject-based community exchanges), we can also observe domain-specific terms, high lexical coherence, more levels of syntactic embedding, more complex readability levels. These are in fact, in our view, the hallmarks of knowledge building and informative flow.

#### **4. Concluding remarks**

NLP tools for content analysis and Information Extraction are instrumental in establishing 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. As cognitive proximity consists of sharing capabilities and knowledge in a broad context, subject based communities, primarily focused on supporting relationships and content sharing, act at the same time as providers and users of all kind of Grey Literature materials in a highly distributed and collaborative scenario, and represent a conducive environment for knowledge creation and transfer. They can represent, as collaborative networks, a key element in the advancement and dissemination of knowledge in scientific domains as well as in diverse aspects of everyday human life.

Conversely, general-purpose social networks, reflecting either friendship or superficial relationships, are virtual meeting place, but tend to generate ephemeral information and to create superficial and mosaic knowledge.

As a general concluding consideration, Social Networking is a medium with a strong potential, a house of cards powerful and delicate at the same time.

## References

- ANDERSON S. R., LIGHTFOOT D. W. (2002). *The language organ: linguistics as cognitive physiology*. Cambridge University Press.
- BONIN F., DELL'ORLETTA F., VENTURI G., MONTEMAGNI S. (2010). A Contrastive Approach to Multi-word Term Extraction from Domain Corpora. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Malta, 19-21 May, 3222-3229.
- BAEZA-YATES R., RIBEIRO-NETO B. (1999). *Modern Information Retrieval*. Addison Wesley, ACM Press, New York.
- BUITELAAR P., CIMIANO P., MAGNINI B. (2005). *Ontology learning from text*. IOS Press, Amsterdam.
- CHURCH, K., HANKS P. (1989). Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.
- DELL'ORLETTA F., MONTEMAGNI S., VENTURI G. (2011). Read-it: Assessing Readability of Italian texts with a View to Text Simplification. *Proceedings of the 2<sup>nd</sup> Workshop on Speech and Language processing for Assistive Technologies*, Edinburgh, UK, 73-83.
- DELL'ORLETTA F. (2009). Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09*, Reggio Emilia, December 2009.
- FRANTZI K. T., ANANIADOU S., MIMA H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3( 2), 115-130.
- LENCI A., MONTEMAGNI S., PIRRELLI V. (2006). Acquiring and Representing Meaning: Computational Perspectives. In A. Lenci, S. Montemagni, V. Pirrelli (eds.) *Acquisition and Representation of Word Meaning. Theoretical and computational perspectives*. *Linguistica Computazionale*, XXII-XXIII, IEPI, Pisa-Roma, 19-66.
- LENCI A., MONTEMAGNI S., PIRRELLI V., VENTURI G. (2008). Ontology learning from Italian legal texts, in J. Breuker, P. Casanovas, M. C.A. Klein, E. Francesconi (eds.), *Law, Ontologies and the Semantic Web - Channelling the Legal Information Flood, Frontiers in Artificial Intelligence and Applications*, Springer, Volume 188, 75-94.
- MANNING C. D., SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- MARZI C., PARDELLI G., SASSI M. (2011). A Terminology based re-definition of Grey Literature. *GL12 Conference Proceedings*, TextRelease, Amsterdam, 27-31.
- MARZI C., PARDELLI G., SASSI M. (2010). Grey Literature and Computational Linguistics: from Paper to Net. *GL11 Conference Proceedings*, TextRelease, Amsterdam, 118-121.
- NOOTEBOOM B. (2000). *Learning and innovation in organizations and economics*. Oxford University Press, Oxford.
- SMADJA, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, MIT, Cambridge MA, USA, 19(1), 143-177.
- TENENBAUM J. M. (2006). AI meets Web 2.0 Building the Web of Tomorrow, Today. *AI Magazine*, American Association for Artificial Intelligence, 47-68.
- TEXT ANALYSIS TOOLS <http://www.ilc.cnr.it/dylanlab/> developed at *Dylan Lab* (ILC-CNR).