INNOVATION, LANGUAGE, AND THE WEB

Claudia Marzi
Institute for Computational Linguistics "*Antonio Zampolli*" *- National Research Council, Italy* (*CNR*)
University of Pavia – Dept. of Theoretical and Applied Linguistics
Via G. Moruzzi 1, 56124 Pisa, Italia
claudia.marzi@ilc.cnr.it

Abstract

Language and innovation are inseparable. Language conveys ideas which are essential in innovation, establishes the most immediate connections with our conceptualisation of the outside world, and provides the building blocks for communication.

Every linguistic choice is necessarily meaningful, and it involves the parallel construction of form and meaning. From this perspective, language is a dynamic knowledge construction process.

In this article, emphasis will be laid on investigating how words are used to describe innovation, and how innovation topics can influence word usage and collocational behaviour.

The lexical representation of innovative knowledge in a context-based approach is closely related to the representation of knowledge itself, and gives the opportunity to reduce the gap between knowledge representation and knowledge understanding.

This will bring into focus the dynamic interplay between lexical creativity and innovative pragmatic contexts, and the necessity for a dynamic semantic shift from context-driven vagueness to domain-driven specialisation.

Topics: Research life circle; New Technologies; Linking research.
Keywords: *Lexical productivity, Language technologies, Web corpora, Grey Literature.*

1. Introduction

Understanding the relationship between language and innovation is connected with understanding that language determines what can and what cannot be talked about, and therefore what can be achieved and what cannot.

Language conveys innovative ideas and gives body to knowledge itself. Language is the common ability of our species that makes us shaping and referring to things, events, and concepts with

remarkable precision. This common communicative ability connects people into an information-sharing network, and information allows people to expand their knowledge.

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.

Knowledge transfer and innovation transfer are nowadays ubiquitous processes. The entire system of knowledge refers to knowledge creation and application, by defining a process going form acquisition and sharing to transfer and application. Knowledge extraction involves heterogeneous tasks related to the acquisition, from unstructured textual data in digital format, of structured and classified information relating to research topics. Nonetheless, far from being readily or easily transferred from the originator to the user of a technology, knowledge faces barriers, such as ambiguity, difficulty to be interpreted and absorbed, difficulty to be retrieved.

The spread of Internet has enabled development of better bibliographic scientific databases with significantly improved capacity for storage and retrieval. In recent years, web searching has become the default mode of highly innovative information retrieval, though the main sources of digital information are unstructured or semi-structured documents. Information relating to developments in scientific research is collected in the form of abstracts or full publications, in large and growing bibliographic repositories.

Knowledge ambiguity may also depend on language ambiguity; a shared language is part of a related language, and “for knowledge to be exchanged and combined, there has to be a shared medium of communication” (Hedlund, 1999:11). Language has effects at all stages of knowledge transfer. Language generates on-going impacts beyond a simple knowledge transfer act, and it is simultaneously an active agent in the knowledge transfer process itself.

Our goal is to focus on how words and language structures become vehicle for knowledge generation, and in particular for innovation transfer, and what kind of infrastructural support can enhance innovative knowledge transfer.

2. Background

2.1. Language of innovation and linguistic innovation

In considering the language of innovation, particular emphasis is laid on overlap and dissonance in terminology and word formation processes associated with innovation. The language of innovation suffers from a problem of lexical overabundance. Not only many words are offered, but different authors define or use these words in different ways; and this because the challenge is the complexity of categorising innovation.

Many terms are used to describe innovation (Linton, 2009, for an overview), and innovation itself offers new collocations and extension of use. Most of the differences in terminology can be accounted for by differences in perspective and domain.

Language change is a fundamental evolutionary phenomenon due to many natural, cultural and historical factors. In particular, we are interested here in shedding light on the phenomenon of terminology innovation and propagation of those novel forms across domains.

Through language cultural novelty can be transmitted vertically (from parents to children), horizontally (from peer to peer) as well as across generational gaps. Although many linguistic innovations fall into the category of what Andersen (1989) has called “fortuitous innovations” (i.e. spontaneous and purposeless innovations as the results of non-functional, non-intentional copying errors), in many cases the very structure of the innovation can be explained with reference to the speakers’ (synchronic) perception and meaningful re-interpretation of linguistic surface forms within the pragmatics of the situation. Meaning itself is a consequence of interaction and context.

The meaning of words can change over time and discourse and, in particular, words can take on new senses when used in novel contexts. Words with emergent novel senses often reflect an extension of use from one domain to another. In this sense, linguistic innovations arise in the context of existing rules which they modify.

2.2. Language ambiguity

Many words have more than one sense, and each of their sense is reflected by their distribution across contexts. Lexical semanticists make a classical distinction between semantically ambiguous words (e.g. *bank*), whose meanings can vary unsystematically, and polysemous words (e.g. *school*), where different senses exhibit a predictable relationship. Both ambiguous words and polysemous words can be disambiguated by defining their context of use. Polysemous words, in

particular, can shape their meaning as a function of their context of use. As a consequence of this context-sensitivity, if a polysemous word-type appears more times in the same text (e.g. a single document), it is extremely likely that its different tokens will share the same sense. Although people do not need too much context to perform a disambiguation task, in Natural Language Processing (NLP) and Information Retrieval (IR) larger contexts make the task easier. Methods for computing relational similarities and disambiguating polysemous words, based on large text corpora, can make rough sense distinctions. Although this is far from reaching the sophistication of human judgement, the field is making considerable progress in context-sensitive word sense disambiguation and in the identification of conceptual relations between words.

In the following sections, corpus-based investigations are analysed in the perspective of a lexical representation of innovative knowledge. In the field of corpus linguistics advantages and limits of various corpora are analysed, depending on both linguistic and innovative knowledge research questions.

2.3. Corpus linguistics

Corpus-based linguistics is the study of language on the basis of large text samples – the corpora.

A corpus, as defined by Sinclair (1999: 171), “is a collection of naturally occurring language text, chosen to characterize a state of variety of a language. In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are clues to the lexical structure of the language”. Although it must be considered that an appropriate size of corpus is strongly dependent on the phenomenon to investigate and the purpose itself. Another factor influencing the size of corpora relates to the degree of internal variation in the language or genre under study (Meyer, 2002).

In any case, corpora are incomplete. Rather, the issue is whether they are representative of the inquired phenomena; in other words, a corpus should be large enough to give an adequate representation of the language and more occurrences of the elements under investigation.

The importance of findings, either quantitative or qualitative, depends on the representativeness of the selected corpus for the research question.

Dealing with machine-readable texts offers the basis for purpose-specific research questions.

Occurrence, distribution, and importance, are different issues to be taken into account.

Salient domain-specific concepts and relations are most often conveyed in text through statistically significant terms. Rare words often denote the most salient pieces of content information of a document together with its level of subject-specificity. Recurrent word combinations are defined as COLLOCATIONS.

In corpus linguistics, collocation defines a sequence of words or terms that co-occur more often than would be expected by chance. An example of a phraseological (multi-word expression) collocation is the expression *strong tea*. While the same meaning could be conveyed by the roughly equivalent **powerful tea*, this expression is considered incorrect by English speakers. Conversely, the corresponding expression for computer, *powerful computers* is preferred over **strong computers*. Unlike idioms, collocations have a rather transparent meaning and are easy to decode; yet they are difficult to encode – like idioms – since they are unpredictable for non-native speakers, and moreover they do not preserve the meaning of all their components across languages. Collocations are flexible and they can involve two, three or more words in various ways.

In NLP collocational information derived from corpora is useful in the perspective of text analysis; for instance, in word sense disambiguation collocations are used to discriminate between senses 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. Corpus-based studies on word formation show that productivity of derivational suffixes are more pertinent in certain kinds of texts than others. In other words, register variation plays a salient role in word formation. It's, however, essential to state that register distinctions are not defined in linguistic terms, but rest on context, domain, and purpose; and that contextual knowledge allows to support knowledge processes and to better access them.

2.4. The Web as a corpus

The World Wide Web has become a primary meeting place for information and communication, and it provides texts to be mined for lexicographers and linguists. As the web is constantly

expanding, it represents an unlimited universe of information and data, and offers ubiquitous accessible information, and large volumes of information are available. Increasingly, corpus linguists have begun using the World Wide Web as a corpus for linguistic analyses.

The Web as a linguistic corpus makes it possible 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 for corpus linguists and researchers in the fields of NLP, IR and Text Mining, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. The Web and associated technologies have been both the catalyst for much linguistic creativity and the main vehicle for its dissemination. In contrast, any static corpus is cut off at the moment of its compilation.

However, the web is a particular kind of corpus, as an estimation of its size and especially its composition cannot always be assessed. Moreover, it must be seriously kept into consideration that investigation of selective corpora is better concerned with the description of use and structure of domain-specific language, by inquiring linguistic phenomena, such as co-occurrence distributions, collocational variability, derivational productivity, neologism coinage. While the notion of linguistic corpus as a body of texts rests on some related issues such as finite size, balance, permanence, the very idea of a web of texts brings about notions not only of flexibility but even of non-finiteness and provisionality.

How can data gathered from the web provide new insight into language usage? The Web as a corpus is a rich source of freely available linguistic data covering a lot of topics (Fellbaum 2005), but despite the great advantages in quantity and accessibility, there is no control for example on web posting, and especially concerning English data, a lot of data are posted by non-native speakers. In this sense, the language used on the Web does not represent thoroughly the standard usage. While statistically robust analysis of Web data to discover collocations can give a flavour of what Manning and Schütze (1999) defined as “a conventional way of saying things” by marking the most frequent expressions, high frequent occurrences cannot give a disambiguation of context usage and sense. Moreover, in using the web as a corpus especially when it is accessed through generic search engine, it is virtually impossible to replicate a test on the same data.

In short, the Web offers a huge repository of documents written in a multitude of – more or less standard – languages, of different types or genre, and constantly changing over time, though often helpless in telling intended purposes and in offering background and contextual knowledge.

In what follows we propose to approach the tight inter-relationship between sense extension, context and innovation, by exploring the usage of words in contexts with NLP technologies. In Computational Linguistics and Computational Lexicology, sense identification and words sense disambiguation are commonly modelled by focusing on the distributional similarity of word usages in context. These techniques provide a key to a deeper understanding of a constructive view of lexical meaning as the by-product of the interaction of a word with its surrounding context (i.e. its collocates), and represent the methodological basis of the ensuing analysis.

3. Methodology and experimental evidence

3.1. Method and materials

The challenge of identifying changes in word sense has only recently been considered in Computational Linguistics.

To investigate the themes discussed in the previous sections genre-oriented and stylistically heterogeneous English texts are analysed, with the support of SKETCH ENGINE (Kilgarriff et al., 2004), which is a corpus query tool, based on a distributed infrastructure, that generates *word sketches* and *thesauri* which specify similarities and differences between near-synonyms. By selecting a collocate of interest in a *sketched* word, the user is taken to a concordance of the corpus evidence giving rise to that collocate.

Ambiguous and polysemous words have been selected with particular reference to innovative domains, and their collocations are analysed. In particular, we considered the domain of brain sciences and new technologies of brain functional imaging, the domain of knowledge management processes, and the field of information technologies, by mainly focusing on the following test words: IMAGING, RETENTION, STORAGE, CORPUS, NETWORK, GRID.

The selected words present a potentially high degree of semantic ambiguity or polysemy and different degrees of semantic specialisation, which can be analysed objectively by studying their context collocations.

For a terminology exploration, both domain-specific and general-purpose texts materials are selected by using generic search web engine queries (www.google.com by using seed words), domain-specific databases and type coherent multidisciplinary large corpora (e.g. www.opengrey.eu, www.ncbi.nlm.nih.gov/pubmed by selecting the domain). Collocations and concordances are then compared with large balanced corpora (e.g. the British National Corpus, British Academic Written English, New Model Corpus, and the like, whose size ranges between 8 M and 12 G tokens).

3.2. Results

By comparing in different contexts of use collocates and keywords selected from reference corpora – both specific and generic – with simple keywords search in web context, we investigate the ambiguity vs. polysemy gradient, showing how dynamically word meanings are adjusted to novel usage, and how difficult it could be to disambiguate polysemy words without a predefinition of the specific context.

All six terms exhibit distinct senses when they are used in different domains/discourse contexts. However, the extent to which different senses of the same term are mutually related differs considerably from one term to another. The two different usages of Latinate CORPUS as referring to the medical domain, and in particular to brain areas (e.g. CORPUS CALLOSUM, CORPUS STRIATUM) as opposed to large collection of items in a general sense and to large collections of texts/specimens in the Humanities are related only etymologically, with no systematic sense shift or extension. Moreover, the use of CORPUS as 'collection' is by far more widely-spread than its (neuro)-anatomical sense. If we are not in a position to constrain automatic word search within particular text domains, the medical usage of CORPUS is likely to be severely under-represented, flooded by the vast majority of examples of the more generic sense.

In the case of IMAGING on the other hand, the prevalent use of the polysemous term in connection with the medical domain, as referred to specific diagnostic technology, is the result of the application of a general-purpose technology to a specific domain. Technical and scientific bibliographic databases present only this collocate, while by selecting it as a seed word in a web engine search, the more frequent collocation is the one referred to generic visual representations.

The selected balanced corpora, on the other hand, exhibit both of them, with a higher frequency of occurrences in the medical domain, in particular in the brain sciences.

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. NETWORK and GRID are expression of the very popular domain of information technology, and even though related to innovation, they are identified as related to this specific context, exhibiting a very coherent collocation behaviour.

Finally, RETENTION and STORAGE appear to oscillate between their proper and extended senses interchangeably, thus witnessing a paradigmatic case of systematic, context-sensitive polysemy.

RETENTION can select both material and immaterial items, but in the bibliographical references system on Grey Literature presents collocations in the domain of information and knowledge retention, as part of the whole process of acquisition, storage and retrieval, whereas in biomedical scientific database is specialised for fluid retention. STORAGE can make reference to containing units, either physical or computational, to long term memory capacity, related to either computational or cognitive processing. Moreover, the storage process is part of the above mentioned knowledge process. Thus, reference to the storage process by itself does not disambiguate the object to be stored.

In figure 1, logarithmic relative frequency distributions across domains are plotted for the test words. Firstly, for each of the selected words, frequencies – expressed as a relative percentage of occurrences in a balanced corpus - are taken and ranked for the diverse domains. Secondly, the top ranked domains are selected and presented together.

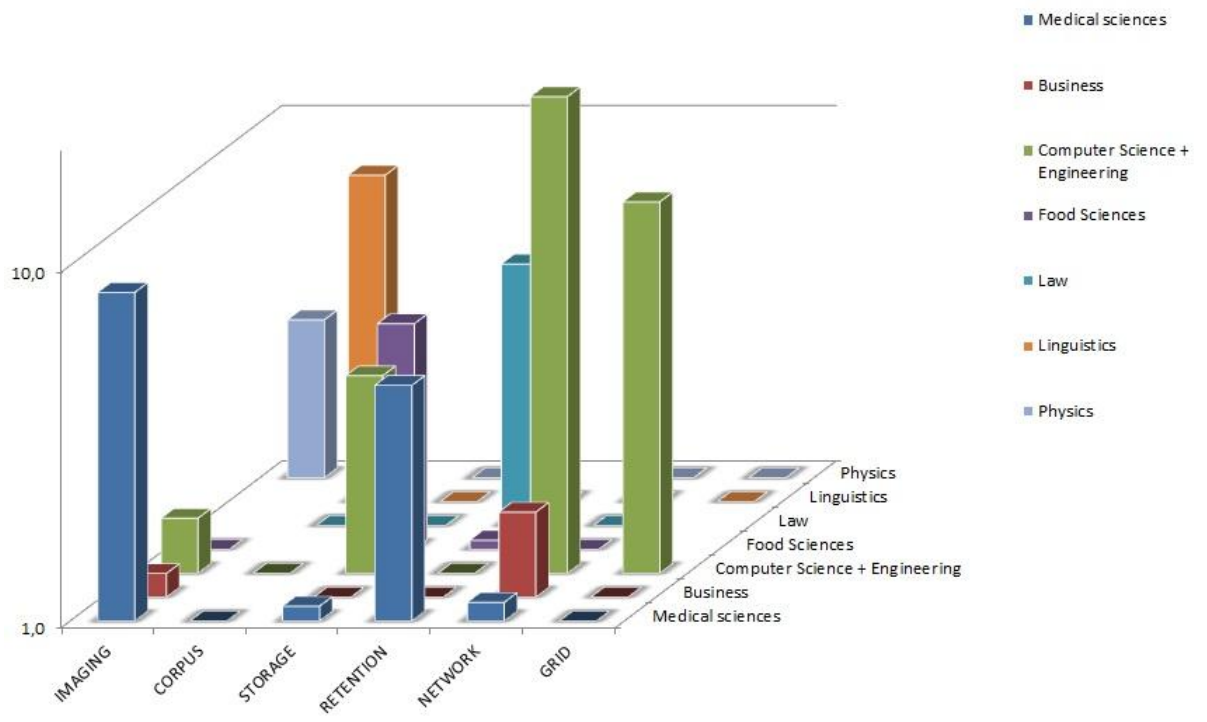
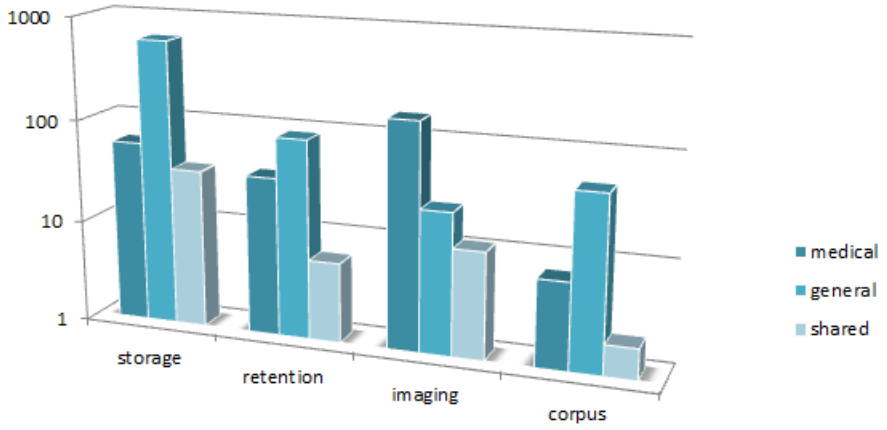


Figure 1- Test words log relative frequency distributions across domains.

Figure 1 provides a snapshot of the usage of our test words across domains, showing that some words are chiefly used in some domains only; however, it does not lay emphasis on the fact that all six terms exhibit multiple senses but relate them differently. The inter-sense relationship can be described in terms of i) AMBIGUITY, when multiple senses show no common schema/relation (as in the case of CORPUS), ii) POLYSEMY, when senses are specialisations/extensions of core meanings (see for example IMAGING); and iii) VAGUENESS, when words can interchangeably be used in both generic and domain-specific contexts with no appreciable sense shift (as with RETENTION and STORAGE). There are words (NETWORK and GRID) which show domain-sensitive specialisation and are chiefly used in their specialised sense.

To further investigate the relationship between different senses, we measured to what extent words which are used both in a specialised context and in a general context tend to co-occur with the same collocates, i.e. tend to be used in similar contexts. This is illustrated in Figure 2, where, for each test word, we counted the number of top-ranked collocates for each of two domains (medical vs. general), and then the number of collocates present in both ranks. As expected, the

term CORPUS exhibits the lowest number of common collocates, due to the fact that the two senses of the word share no common meaning core.



	storage	retention	imaging	corpus
medical	59	35	159	7
general	603	87	24	49
shared	35	6	11	2

Figure 2 – Medical sciences and general domain shared collocates

It is important to note, however, that words like STORAGE, RETENTION and IMAGING present more overlapping contexts and different degrees of domain specialisation. In particular, IMAGING is by far the most frequent term in its medical usage, even in a non-specialised domain, as shown by the considerable number of collocates that are common to both medical and general domains. Moreover, it is often very difficult to distinguish two different senses of the same term on the basis of the observation of its contextual behaviour only. It turns out that terminological innovation – as expressing conceptual innovation - often implies a gradient extension of a core meaning rather that a radically different usage. We doubt that distributional analyses of contexts can provide a fully automatic basis to tackle sense distinctions at this level of granularity and flexibility. Nonetheless, innovation and domain-driven specialisation can be characterised distributionally in terms of a correlation between typical contexts of usage and knowledge domains.

To sum up, IMAGING, RETENTION, STORAGE, and CORPUS versus NETWORK and GRID are examples of linguistic knowledge in a very specific context-based approach, and therefore examples of phenomena such as ambiguity and polysemy. They are closely related to a representation of innovative knowledge, though still representing an existing gap between the semantic, context and

domain dimensions. On the other hand, the relationship between meanings and reality, and lexical semantics are bridged in linguistic data related to concepts of the field of information technology, which is widely known, identified, spread and popular so to enhance a common and coherent understanding.

4. Discussion and concluding remarks

The spread of internet has enhanced the development of better bibliographic scientific databases with a significantly improved capacity for storage, access and retrieval. In recent years, automated search of electronic database has become the default mode of scientific information retrieval.

As an answer to the need for having better domain-specific databases available for improved storage and retrieval of scientific content, bibliographic domain-specific and type-oriented databases are developed, and emerged over the last two decades to offer materials to a specialised readership, thus providing highly-selected, well-targeted documents.

Such specific infrastructures define a communication network, where writers and readers – like speakers – are in a sufficient proximity to each other to have very high probability of communication with each other, by sharing the same “language", intended as the population of utterances in a speech community (Croft, 2000). The informative purposes are defined by the context, in line with Jakobson’s model of communication (1960), where six constituents of the communication act are modelled as functional roles: the addresser or encoder, a message or a signifier, the addressee or decoder, a context or the signified – where by context Jakobson means referent, a code or shared mode of discourse, and a contact or channel.

Analysis of word usages in large corpora offers the opportunity to investigate how words and language structures become vehicle for innovative knowledge generation and transfer. Lexical co-occurrences and collocations can be of considerable help in retrieving text materials which are relevant to a specific domain of interest, but they can be of little help in distinguishing one particular sense of a polysemous word from its other senses. This is especially true of innovative usages of existing terms, which appear to transfer and adapt their original and more general collocates to one or more specialised domains. This is the reason why automated sense disambiguation based on

the distributional analysis of words in context proves to be less effective especially in distinguishing word usages which are most strongly related to innovation.

Due to ambiguity, polysemy and homography, most terms are multi-referential or multi-contextual (Renouf, 1993) in use. Low-precision results could be improved by restricting the contextual domain. A more supervised approach to the problem, where domains and texts are classified preliminarily by domain experts, rather than being bootstrapped from patterns of word distribution only, promises to be more successful in this task.

We investigated the hypothesis that extension of usage process and polysemous disambiguation correlate significantly with genre- and domain- oriented texts and intended readership, thus providing a convenient way to track down well-targeted, highly technical repositories of openly available text materials.

In requiring a dynamic shift from context-driven vagueness – in term of semantic polymorphism - to domain-driven specialisation – in term of terminological usage, the lexical representation of innovative knowledge in a context-based approach is closely related to the representation of knowledge itself, and represents the opportunity to reduce the gap between knowledge representation and knowledge understanding.

Our main emphasis is laid on the importance of effective information technology repositories and distributed infrastructures for the implementation of knowledge processes, and for an efficiently and widely distributed dissemination of research and innovation results, so to enhance future research.

References

ANDERSEN H. (1989). Understanding linguistics innovations. In Breivik, L.E., Jahr, E.H. (eds.), *Language Change: Contributions to the Study of its Causes.* Mouton de Gruyter, Berlin. 5-27.

BAEZA-YATES R., RIBEIRO-NETO B. (1999). Modern Information Retrieval. Addison Wesley, ACM Press, New York.

BIBER D. (1989). A typology of English text. *Linguistics,* 27, 3-43.

BUITELAAR P., CIMIANO P., MAGNINI B. (2005). Ontology learning form 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.

CROFT W. (2000). Explaining Language Change. An Evolutionary Approach. Longman, London.

DEUMERT A. (2003). Bringing speakers back in? Epistemological reflections on speaker-oriented explanations of language change. *Language Sciences*, 25, 15-76.

FELLBAUM C. (2005). Examining the constraint on the benefactive alternation by using the World Wide Web as a corpus. In Kepser S., Reis M. (eds.) *Linguistic evidence: empirical, theoretical, and computational perspectives.* Walter de Gruyter, Berlin. 209-238.

FLETCHER W.H. (2011). Corpus Analysis of the World Wide Web. In *Encyclopedia of Applied Linguistics.* Wiley-Blackwell.

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.

HEDLUND G. (1999). The Multinational Corporation as a Neatly Recomposable System. *Management International Review* 39,1. 5-44.

JAKOBSON R. (1960). Linguistics and Poetics: Closing Statement. *Style in Language*. MIT Press, Cambridge, MA.

KILGARRIFF A., RYCHLY P., SMRZ P., TUGWELL D. (2004) The Sketch Engine. *Proceedings EURALEX* 2004, LORIENT, FRANCE. 105-116.

LINTON J. (2009). De-babelizing the language of innovation. *Technovation*, 29, 729-737.

LÜDELING A., EVERT S., BARONI M. (2006). Using web data for linguistic purposes. In Hundt M., Nesselhauf N., Biewer C. (eds.) *Language and Computers, Corpus Linguistics and the Web* 18, 7-24.

MANNING C. D., SCHÜTZE H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

MARZI C. (2012). Knowledge Communities in Grey. *Proceedings of the GL13 – International conference on Grey Literature: the Grey Circuit.* TextRelease, Amsterdam. 34-40.

MEYER C.F. (2002). English Corpus Linguistics: An Introduction. Cambridge University Press, Cambridge.

NOOTEBOOM B. (2000). Learning and innovation in organizations and economics. Oxford University Press, Oxford.

PANGARO P. (2008). Innovation, Language, and Organizations. *Continuum Itaú Cultural magazine*. 7.

SERETAN, V. (2011). Syntax-Based Collocation Extraction. Springer, Heidelberg-London-New York.

SINCLAIR J. (1991). Corpus, concordance, collocation: Describing English language. Oxford University Press, Oxford.

SMADJA, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, MIT, Cambridge MA, USA, 19(1), 143–177.

STUBBS, M. (2001). On inference theories and code theories: Corpus evidence for semantic schemas. *Text*, 21 (3), 437-465.

RENOUF, A. (1993). What the Linguist has to say to the Information Scientist. In Forbes, G. (ed.) *The Journal of Document and Text Management* vol. 1/2, 173-190.

VETERE G., OLTRAMARI A., CHIARI I., JEZEK E., VIEU L., ZANZOTTO F.M. (2011). Senso Comune, an Open Knowledge Base for Italian. *TAL*, 52 (3), 217-243.

WELCH D., WELCH L. (2008). The importance of Language in International Knowledge Transfer. *MIR*, 48(3), 339-360.

THE SKETCH ENGINE http://www.sketchengine.co.uk