THE DEVELOPMENT OF A CORPUS-BASED LEARNING OBJECT TO IMPROVE WRITTEN ENGLISH LANGUAGE SKILLS IN ITALIAN COMPUTER SCIENCE PHD STUDENTS

Lynn Rudd and Antonietta Bagnardi
(*University of Bari, Italy*)

Abstract

The decision to develop a specific corpus-based language learning object was based on the need of the PhD students in the Department of Computer Science at the University of Bari (Italy) to enhance their English language abilities for writing theses and articles. This need was clearly recognized in the Department both by the discipline specialists and by the authors – the ESP language specialists – thanks to their vast experience in reading and correcting such manuscripts.

A learning object was considered the best, most innovative method for transmitting self-contained "chunks" of linguistic information to busy PhD students. In order to establish a suitable approach to the development of the learning object, the authors performed an in-depth study of the literature concerning academic, scientific and specialized discourse analysis and works regarding corpus analysis. The authors also considered it important to make comparisons between the learner (L1) 'corpora' and the native speaker (L2) 'corpora', in order to help solve problems of non-native speaker writers of academic English texts.

An accurate analysis of the students' starting level and linguistic needs was then performed, by means of a written test. To build the ESP corpus, a selection of specialized academic writings, published in Computer Science journals, was collected, chosen by the discipline specialists. These target papers were presented to the students and each section was analysed both stylistically and grammatically, concentrating on the recurrent lexical and syntactic features of the micro-language used in Computer Science (for example, the common use of acronyms, nominal groups and specific phrasal verbs). Hence the learning object gives the students a basic model of how to build a paper/thesis in the specific field of Computer Science, through the identification of these key corpus tools. The learning object is described in detail in the core of this paper, including the guidelines, hints and useful activities provided to aid the learning process. The results of the student training are presented and some conclusions are drawn regarding the efficacy of the approach adopted and the suitability of the corpus size.

1. Introduction and background

Computer Science PhD students need to acquire specific skills in English if they want to reach a better performance level in that language. These skills are 1) receptive:

reading and understanding vast quantities of specific material in English (e.g. papers in journals, reviews and theses – taken from the Internet and libraries); 2) productive: PhD students are required to participate in writing papers with their professors and to prepare their own theses in clear, concise English, demonstrating not only an in-depth knowledge of the scientific/technical concepts, but also a good command of the micro-language used in their specific field of study. They may also be required to present their work orally to an audience (at conferences or when discussing their theses) in coherent, comprehensible English. However, due to specific indications from the discipline specialists and the authors' vast experience of correcting the PhD students' written work, this particular study will deal exclusively with improving English writing skills, to meet the most immediate needs of these students.

We considered a learning object to be the best, most innovative means for a rapid transmission of self-contained "chunks" of linguistic information to busy PhD students, in order to improve the required techniques. In fact, a learning object has been defined by Beck (2008: 64) as "a collection of content items, practice items, and assessment items that are combined based on a single learning objective." Learning objects are known by many names such as "content objects", "chunks", "nuggets", "knowledge bits", "media objects" and "testable reusable units of cognition". Beck affirms that learning objects offer a new way of dealing with the learning process. Instead of the traditional "several hour chunk", they provide smaller, self-contained, reusable units of learning. However, as stated by Kendall, Wakefield and Delbridge (2007: 55), they "must be large enough to have educational value". Therefore, we chose a small-medium-sized learning object to facilitate the learning process of the PhD students, who have to intersperse their heavy study load in their subject matter with precious "nuggets" of linguistic information.

This choice was also based on our previous experience in the development of an e-learning object for the teaching of English in the Bachelor's degree course in Computer Science (Bagnardi and Rudd 2012). The development of a small-medium-sized e-learning object makes it "communicative, user-friendly and enjoyable to use." We realized that a learning object functions like a computer processor, which considers and selects the input, analyses and processes all the information received and yields a specific output. In the particular case described in this study, the input will be the specialized Computer Science target texts that form the corpus, which will be analysed and processed by the PhD students. The desired result will be to see these novice writers building their own models and creating well-structured and linguistically rich papers, to make successful written contributions in their new community of practice.

In order to establish a suitable approach to the development of the learning object, we carried out an in-depth study of the literature regarding academic, scientific and specialized discourse analysis and works concerning corpus-based language learning.

Recent analyses have been carried out on the academic language produced by native speaker writers, compared to that produced by non-native speaker writers. Gilquin *et al.* (2007) stated that novice non-native speaker writers of academic prose have similar problems to novice native speaker writers, but they also have their own specific difficulties. They maintain that it is not sufficient to use information only from a native speaker corpus (L2); a study of the learner corpora (L1) is also required to be able to understand and to try to solve the problems related to the lexico-grammatical patterning of non-native speaker writers of academic English texts. This reflection induces us to

consider carefully the similarities and differences between L1 and L2 in our learning context and to underline the specific points that may cause difficulties.

De Saussure (2002: 272) defines 'style' related to a common social context, to be studied through analogies and associations: each individual can have his/her own writing style. According to De Saussure, style means studying the means of expression, if language is perceived as a 'common' vehicle of communication. It is not governed by specific rules and only through the observation of other writers' styles may one acquire one's own style. According to Windish (1985), if we want to talk about a 'collective' style, an in-depth knowledge of the functions and the structure of the 'discursive' style of the whole group of learners is required. Moreover, the same reasoning can be applied to grammar, which De Saussure (1986: 185-187) defines as a complex and systematic object that brings together "co-existent values", such as syntax, which is "a theory of syntagms and associations". This leads us to understand that style is specific not only to the individual writer, but to a community of writers with similar academic interests – in our case Computer Science PhD students and specialists.

Sinclair (2004: 275) investigated language as concepts and links which would generate corpora: "A good pedagogical description of a language will organise the variation and prioritise the variants for language teaching purposes". This process of classification moves from the initial listings of variants to an understandable pattern of relationships that can be arranged according to different criteria, such as complexity and familiarity. Here the most superficial findings of corpus analysis can be used – the frequency of occurrence of items.

Contributions by Swales (1990, 1994 and 2004) were also taken into consideration for genre analysis within the specific context of English for Academic Purposes. Swales describes how distinctive linguistic features of academic discourse can help novice writers overcome the difficulties they encounter when producing academic language. He concentrates on genre analysis as a heuristic tool designed to facilitate the understanding of the ever-evolving nature of genres. Moreover, Hyland (2008b: 45) defines word clusters or strings as "building blocks of coherent discourse which span structural units" and views clusters as "a psychological association between words" (*ibid.* 59). These clusters are patterns that readers and writers of a specific language genre learn to anticipate through frequent exposure and usage. We therefore intend to involve our students to the full in the reading and writing of such patterns.

Considering scientific discourse, Tarantino (2004: 70) observes that scientific progress is a "cumulative enterprise which advances through the contributions of a community agreement […], sharing a number of materials and conceptual tools". Indeed, she also notes (Tarantino 2005b: 74-75) that "common ground knowledge" should not remain "anchored to the grammatico-semantic analysis of sentences, but expand into a collaborative exchange of expressions in order to compare, associate and evaluate new input." This shows that there is a common scientific discourse on a large scale; however, each branch has its own micro-language. Tarantino (*ibid.* 119) also affirms that the language of science depends "not just on syntax and the lexicon, but on modes of thoughts and procedures shared by individuals and communities of practice". This indicates that there is far more to scientific discourse than the mere use of specific terminology.

Halliday (1988: 162) defines scientific English as a functional variety of registers: "A register is a cluster of associated features having a greater-than-random (or rather,

greater than predicted by their unconditioned probabilities) tendency to occur". Moreover, he views language as a complex of choices among "mutually exclusive options", i.e. if one makes an association of words based on experience of the language register, certain other words that are not appropriate are automatically excluded. For example, in Computer Science language the expression "neighbouring clusters" automatically excludes the association of "near clusters". This kind of association clearly shows that there is a specific corpus-based language for each branch of scientific discourse.

McEnery, Xiao and Tono (2006: 15) differentiate between general and specialized corpora: "General corpora typically serve as a basis for an overall description of a language or a language variety. In contrast, specialized corpora tend to be domain or genre specific." Moreover, McEnery and Xiao (2010: 364-380) observe that a lexical approach to language is vitally important. They agree with other linguists such as Bahns (1993), Zhang (1993) and Hoey (2000, 2004), that "collocational knowledge" is essential to develop L1/L2 language skills: that is, "teaching habitual co-occurrences of lexical terms" (McEnery and Xiao 2010: 9-10) facilitates the language learning process. We note that in the specific context of Computer Science, for example, patterns of specific verbs related to certain nouns are frequent: "to perform calculations", "to access a database/the Internet", "to display data" and "to store information".

When examining specialized discourse, Gotti (2008: 25) maintains that it does not belong to a completely different genre from general English but is a variety of the general language system. It adopts certain lexical and syntactic features from general everyday discourse and uses them frequently in a specific way. He also states that within specialized discourse itself there are many other varieties: "Just as general language is not a uniform entity but contains many varieties, common rules and features of specialized discourse co-exist with specific ones, separating each variety from the others." This confirms that Computer Science discourse has some features that are common to scientific discourse in general, but at the same time it presents some specific linguistic patterns that distinguish it from any other type of scientific discourse. Gotti (*ibid*. 18-19) also points out that "there is far more than a straightforward lexical distinction at the root of specialized discourse." Register analysis has moved the research focus away from a mainly "statistical-quantitative" approach (the study of the frequency of occurrence of certain lexical items, aided these days by digital concordancing) towards a principally qualitative approach "which seeks to identify peculiarities of specialized texts in a perspective that is not only micro-linguistic but takes into account the discourse in which they are embedded." This reinforces our decision to analyse specialized Computer Science texts in a wide perspective, considering recurrent lexical, syntactic and stylistic features.

Bearing in mind these general and more specific considerations, we decided that, in our case, the learning object should be focused on the analysis and practice of very specialized patterns of linguistic style and features, contained in a small-medium corpus of specific Computer Science discourse. We maintain that the small size of the corpus will accelerate the learning process, by concentrating on the most pertinent micro-linguistic features and patterns as the precious 'nuggets' of essential information to be conveyed to the Computer Science PhD students. Therefore, a corpus-based language learning object seems to be the most suitable means for inputting and processing such data and for outputting positive results.

2. Linguistic needs analysis

Granger, Hung and Petch-Tyson (2002: 245) carefully point out that "[a]pproaches such as Contrastive Interlanguage Analysis (CIA) and Error Analysis (EA)" are fundamentally important in the process of developing a specific learner corpora. Moreover, Gavioli (2006: 115) discusses the importance of the frequency of occurrence of language items and how the teaching syllabus must bear in mind these features in the ESP learning process: "The starting point for these kinds of studies is usually language features that are known to cause perpetual problems to learners." On the basis of these considerations, we decided to analyse the students' most recurrent stylistic and grammatical difficulties in writing academic English in the specific field of Computer Science.

The PhD students' starting level was defined by administering a written test on the use of the recurrent key grammatical structures required for the specific writing skills. They were also asked to produce an introductory paragraph for a thesis or research paper, to see how these structures were put into practice and whether they were able to follow the stylistic code for their area of study. The results of the grammar test revealed that the PhD students' starting level on average was between Level B1 and B2 on the CEFR scale, with one case of C1 level. The average B level was to be expected, as the students had all studied previously for the obligatory English exam during their Bachelor Degree Course in Computer Science. The paragraphs produced revealed some difficulties in the application of certain grammar rules and also some typical stylistic errors.

The two graphs below report the percentages of the most common stylistic and grammatical errors made by the PhD students.

The results of this analysis indicate that stylistic mistakes appear to be less frequent than grammatical errors. This is probably due to the fact that stylistic errors are less

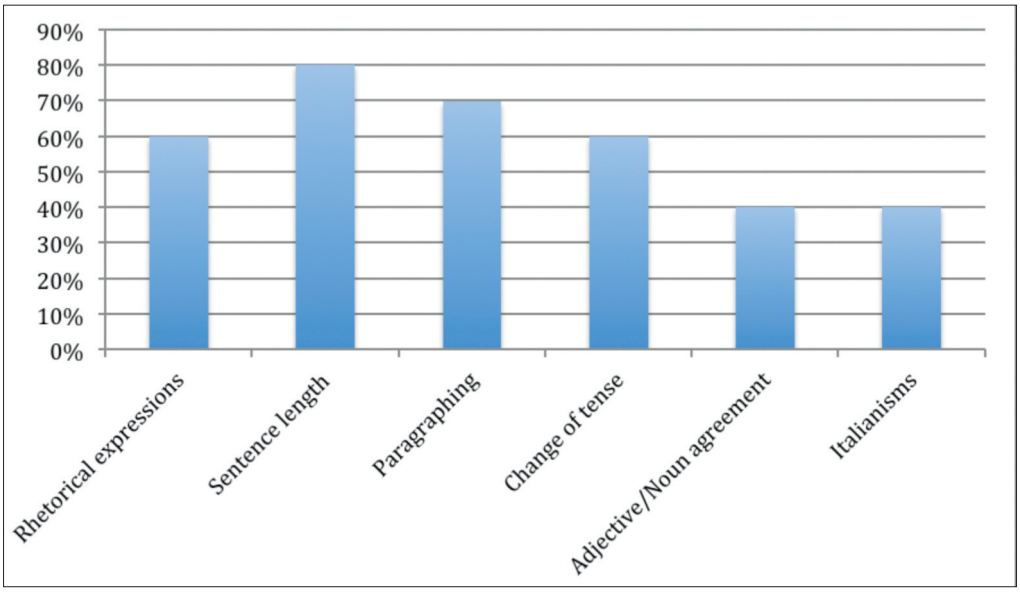

Figure 1. Most recurrent stylistic errors

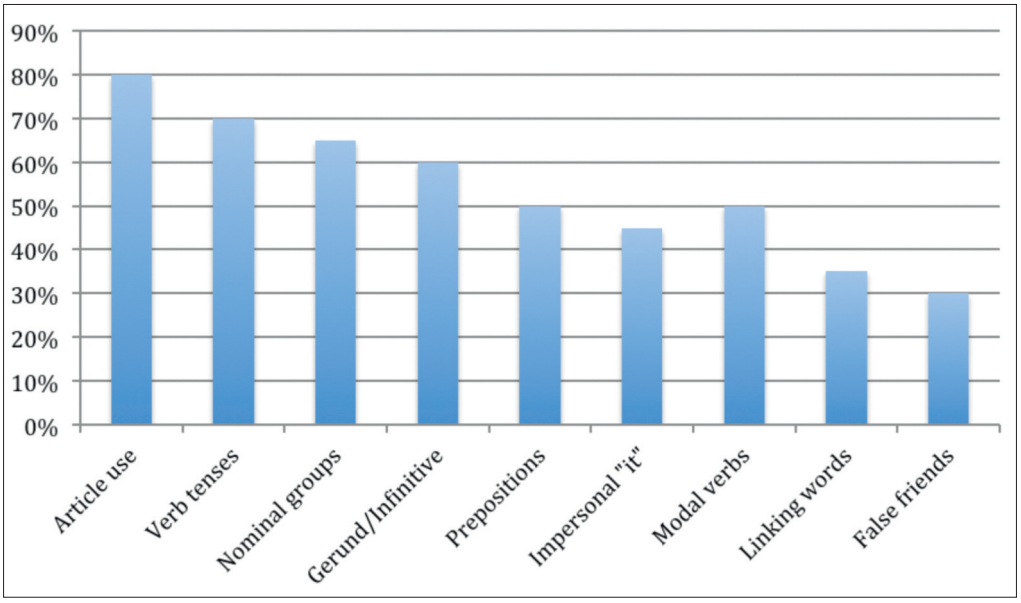

Figure 2. Most recurrent grammatical errors

obvious and are perhaps more difficult to highlight than grammatical errors, but it does not mean they are in actual fact less frequent. Article use and sentence length seem to be the most recurrent mistakes (80%), probably because the Italian language offers different stylistic and grammatical schemata compared to the English language.

On the basis of these observations a series of seminars was organized in which the aim was to provide the PhD students with a model building learning object to solve these potential problems.

3. Methodology and description of the learning object

From a practical, methodological point of view, we considered some experimental works on how to reinforce writing skills in academic scientific English for non-native speakers: Raimes (1999), Wallwork (2011), Glasman-Deal (2010), and Zobel (2014) who focus specifically on academic writing skills for Computer Science. The activities and hints given in these works provided some helpful suggestions for the creation of the learning object; however, we feel that our learning object may be innovative, since it has been designed and tailor-made to fit very specific student needs.

A corpus is a structural set of texts used for linguistic analyses, to check occurrences or validate linguistic rules within a specific language territory. As affirmed by Yoon and Hinvela (2004: 257-258) in language teaching corpora can be considered as a type of foreign language writing aid, as the contextualized grammatical knowledge acquired by non-native language users through exposure to authentic texts in corpora allows learners to grasp the manner of sentence formation in the target language, enabling effective writing. This exposure does not necessarily need to occur through electronic

concordances. In fact, as noted by Willis (1998: 45) the student awareness of the frequency and use of expressions can be suitably enhanced by their own discovery of concordances in authentic texts in the classroom, which can consequently heighten their confidence when using these language features: "Students need to discover and internalize regularities in the language they are studying. If we can place students in the position of researchers this will accomplish these goals neatly and economically."

The criterion on which the learning object was developed was based on the requests of Computer Science professors/researchers of the University of Bari, who noticed many PhD student difficulties when writing a clear, well-ordered paper/thesis in English. With the collaboration of these discipline specialists a small corpus of authentic texts, chosen from Computer Science journals, was created. Some papers were written by English native speakers and some by full professors and researchers in the Department of Computer Science and fully revised by native speakers. As the PhD students were specializing in different fields of Computer Science, we decided to focus on a topic they had all studied during their Bachelor's and Master's degree courses, to facilitate their comprehension of the subject matter: "Data Mining".

The learning object is divided into three parts:

(a) the reading and analysis of the structure of the target papers;
(b) the identification of a specific Computer Science corpus;
(c) the training of the students.

(a) Reading and analysis of the structure of the target papers

The first step was to provide the students with the target academic papers written in English and published in a common field of study: "Data Mining" (see Appendix for the abstracts of the articles chosen). This material was drawn from conference proceedings and journals by the full professors and researchers who follow the PhD students during their studies in the Department of Computer Science at the University of Bari. These target papers would serve as the specialized corpus and the basis for the model building. The students were encouraged to read and analyse the papers via activities developed by the teachers, to foster guided 'discovery learning'.

The students noticed first and foremost the methodical way of thinking demonstrated and the precise order in which these papers were written, mirroring faithfully the technique of Galileo's 'scientific method' (observation-hypotheses-experiments-results-conclusions). It is as if everything were arranged "in mensura et numero et pondere" (Gorni 1990: 87). Subsequently, they discovered the frequent use of definitions, which are concise descriptions of basic properties, objects or concepts. They are very important in science, since they function in a similar way to mathematical formulae: once you understand the stated truths, it is easier to develop and explain new principles. In Computer Science papers, giving definitions is a necessarily recurrent habit, since it helps the readers to understand the interrelationships among various concepts. Moreover, as pointed out to the students, these notions should be well linked by a correct use of suitable connectors also referred to as "mathematical connectives" (Rudd and Butts 2007: 191). This is a crucial step in transitioning from rudimentary to advanced scientific language.

The teachers then guided the students to examine the basic structure of the papers.

They discovered that a paper generally has a globally symmetrical shape, rather like an hour glass: a wide funnel-shaped introduction section and an inverted funnel-shaped conclusions section, where the more general information is conveyed. The funnel-shape narrows for the central methodology and findings sections, where the more specific information is presented and analysed. "Many of the things you need to do in the Introduction are done – in reverse order – in the [...] Conclusion." (Glasman-Deal 2010: 2). The abstract and title were analysed subsequently, as will be explained later. These six sections (title, abstract, introduction, methodology, results, conclusions) are also described in the Computer Science language as "semantic components" which follow a "well-defined order" and are then subdivided into "text regions" or "paragraphs" (Malerba *et al*. 2005). According to Windish (1985: 32) we should not be happy just to analyse words or phrases, that is, discourse analysis is sometimes opposed to lexicology. The discourse should be seen from a wider point of view: in other words, we should not focus just on individual words, but on words as they are used in paragraphs and entire pages.

Paragraph structure was deemed most important. The students were advised to begin a paragraph with a topic sentence that should express the main idea of the paragraph, to which all the other sentences in the paragraph would be linked. The other sentences should define the topic in more detail, elaborate on it and maybe even contradict it. Paragraphs should be neither too short nor too long, to avoid confusing the reader, especially if the topic of the paper is complex. As stated by Glasman-Deal (2010) paragraphs are also a crucial visual aid to effective reading and writing. We are already conditioned to expect a shift or change in topic or type of information when we see the beginning of a new paragraph in a text. When planning the layout of a paper/thesis, it would be useful for the writer to first make a list of the ideas and concepts that he/she wishes to discuss, and then list under each item what he/she would like to say about it. This would provide an effective skeleton structure with well-structured paragraphs that could then be developed into a coherent, logical text.

i) Introduction section

We decided to begin with this section, as writing a good introduction is fundamental in a paper/thesis: it stimulates the reader to proceed and find out more about the results. It is a good idea to start with a striking statement that will create an impact and capture the reader's interest immediately, for example, "Some innovative developments in the field of X could radically change researchers' views".

As proposed by Glasman-Deal (2010), the introduction to a scientific/technical paper should be composed of four main steps:

(a) the definition and importance of the area of research;
(b) some background information and if possible some literature references regarding related work;
(c) a gap in the literature which the author intends to fill with his/her specific study, or a question which he/she intends to examine further;
(d) the author's specific aims and how the paper/thesis will be structured.

This procedure involves starting with more general statements about the area of research and background information, leading to more specific ones about the question

in hand. The students were required to analyse the introduction of a target paper and find each of these steps by labelling the groups of sentences a) – d), as mentioned above. In order to draw the students' attention to the linguistic style used in the introduction, they were then asked to highlight certain rhetorical words and expressions typically used in the different steps: structures such as "is of paramount importance" and "has recently aroused great interest", which are used to establish the importance of the research area generally.

The students were then required to underline the verbs, paying particular attention to the tenses. They noted the typical use of the simple present tense to state known facts.

Sentence length was also dealt with, as the students had a tendency to write sentences that were too long and complex and therefore unsuitable for the specific English discourse genre. We assume that this inclination is due mainly to L1 influence, since Italian academic prose is usually composed of longer, more intricate statements. The students were advised to keep to a concise, simple structure of subject, verb predicate and object complement and to avoid joining what could be separate sentences with a semi-colon, the conjunction 'and', or a relative pronoun. The use of the comma in English texts was noted, since it is often different from the Italian language.

Some typical Italianisms, due to literal translations from the learner (L1) corpora, were also noted: 'in alternative' from the Italian *in alternativa*, instead of 'alternatively'/'on the other hand'; 'differently from', deriving from the Italian *diversamente da*, instead of 'in contrast to'/'contrary to'.

The correct use of certain adjectives with specific nouns was noted, for example, 'This is a paramount phenomenon'. 'Paramount' means 'important', but it is not used to describe a 'phenomenon', rather it is used to qualify 'importance'. More suitable adjectives for qualifying 'phenomenon' would be 'significant' or 'noteworthy'.

ii) Methodology/Procedure section

The methodology/procedure section should contain:

(a) various information and data collected from several sources to justify the choice of the research;
(b) the different steps of the procedure used for carrying out the research/experiment;
(c) the use of specific linking words, like 'first', 'then' and 'finally' to describe the various phases of the research/experiment.

This section should provide details about what was done during the research and is usually described in the simple past tense, using active or passive voice, depending on each situation. However, a recurrent use of the simple present tense has been noted in the Computer Science texts, when the author wishes to make the procedure more immediate for the reader. Verbs such as "propose" and "suggest" are frequently used in the descriptions. In the methodology section the author also makes certain hypotheses, e.g. 'we assume that' and 'we suppose that', as well as the use of adverbs like 'hypothetically' and 'probably', or defines the different stages and issues of the research, e.g. 'we note that', 'we underline that' and 'we point out that'.

The students also realized how important it is to formulate questions of hypothetical condition, such as 'what would happen if'. Let us extract an example from Appice and

Malerba's article "*What would happen if* multiple view clustering *were* considered a solution to the curse of dimensionality problem in classification?" One possible answer to that question is: "*It would probably* generate a single clustering pattern according to all perspectives" (p.12).

During the analysis of the methodology/procedure section, an abundance of linking words was discovered. They were examined and classified according to their specific functions in context: a) giving examples**:** such as (not just 'as'), like**,** for instance, for example (e.g.); b) time sequence and listing**:** first, second, then**,** while, finally/in the end/eventually; c) cause and effect**:** due to/thanks to/owing to, therefore, thus, since/as; d) contrasting: contrary to, on the one hand ... on the other hand, however, whereas/while**,** although; e) defining**:** in other words, that is (i.e.); f) adding information**:** moreover, furthermore**,** as well as**;** g) comparing**:** more than, less than, higher, lower**,** as high as, as low as, more and more, less and less.

iii) Results/Findings section

The students discovered that in Computer Science papers, the results/findings section is often included in the methodology/procedure section and is not a separate section. This shows how the traditional structural pattern of academic papers can often deviate, depending on the trends in specific areas.

This section usually contains:

(a) a comparison with the initial state and the final state of the study;
(b) comments on the findings and the results during the procedure;
(c) a discussion of whether the aims proposed in the introduction have actually been reached.

The simple present tense is mostly used to describe the findings and the facts. The writer is making affirmations of facts that, according to him/her, are now undeniably true and he/she wishes to make the results seem more immediate to the reader. Some use of the present perfect tense was also noted, indicating a very recent action, well linked to the direct present.

Attention was drawn to the most recurrent expressions used in the results section, for example, 'data are collected/gathered', 'an empirical evaluation of the results is carried out', 'the results confirmed that', 'the analysis proves/reveals', 'we observe/consider/discover'. Here the author is clearly conveying crucial observations on the outcome of the procedure/s followed for the study/analysis/experiment conducted.

The use of the gerund or the infinitive after certain recurrent verbs was discovered, e.g. 'We proposed using an algorithm', 'We avoided using that equation to solve the problem', 'This idea is worth being investigated' and 'Jones et al suggested adopting this method'. We noted that the verb 'to allow' can cause confusion for Italian writers, since it requires a direct object before the infinitive in English: 'This method allowed us to collect/the collection of (not allowed to collect) a vast quantity of data'. A particular use of the infinitive with 'to', meaning 'in order to', was noted. Italian writers tend to use the 'for' + infinitive/gerund structure, for example, 'We performed the experiment to prove (not 'for to prove/proving') the theory'.

iv) Conclusion/Discussion section

The Conclusion represents a summary of the results/findings by:

(a) considering the results and findings in their specificity;
(b) making an evaluation and an elaboration of the observations made during the results/findings section;
(c) answering any questions or doubts that may have been left unresolved in the introduction;
(d) defining any possible limitations noticed during the research;
(e) making more general statements about the implications of the results and possibilities for further investigation.

A considerable use of the epistemic verbs has been noted: modal verbs such as 'can', 'should', 'could' and 'may'. In scientific discourse modal verbs are particularly useful in the conclusion/discussion section. They are used when talking about probable or hypothetical situations or when suggesting possible causes or interpretations. Their use is not always easy to determine because every time an evaluation of the conditions and the results should be made. Indeed, the use of the modal verbs depends on the degree of certainty of the statements and the results.

At this stage of the learning object, we felt the necessity of the presence of one of the discipline specialists to clarify some specific notions and concepts in the subject matter. This was done to ascertain that the students had a complete understanding of the specialized corpus terminology.

v) Abstract and Title section

The explanations given by the discipline specialist helped to provide the students with a better general view of the content of the target paper chosen for the writing of the abstract. Once the technical concepts had been well clarified, a possible basic model of the abstract was discussed. The writer:

(a) provides some general background information on the topic, giving some definitions;
(b) combines the method, the general aim and the specific aim;
(c) summarizes the methodology and provides some details;
(d) indicates the achievements of the study and presents the implications.

The abstract should ideally be written after the entire paper/thesis has been drafted. Its specific aim is to summarize the contents of the whole paper, therefore we dealt with its structure after the other sections. In the abstract the topics of each section of the paper should be briefly mentioned. Moreover, a very clear, concise language should be used to help the reader to decide if the rest of the work is relevant to his/her field of studies.

The title is an even briefer summary of the key points of the paper/thesis. It should contain the principal aim of the paper/thesis and the specific area that will be dealt with in the study. The title should contain clear, eye-catching information, so that the reader can judge if the article will be pertinent for his/her own research, just by looking at the title. The habitual use of concise nominal groups to summarize the topic in very few

words can be noted in Computer Science texts, for example, "A Co-training Strategy for Multiple View Clustering in Process Mining".

(b) Analysis of the specialized Computer Science corpus

During the analysis of the target texts a specialized Computer Science language corpus was investigated. The students were guided (i) to analyse the lexicon, (ii) to compare L1 with L2, and (iii) to discover specific Computer Science language tools.

(i) the lexicon

First the different basic lexical categories were identified, such as the article, the noun, the pronoun, the verb, the participle, the adverb, the preposition and the conjunction. Each of these categories was pointed out and analysed in the target texts to underline the importance of these 'parts of speech', or 'word classes' within the specific language corpus. Moreover, this strategy helped to detect the most recurrent errors made by the students and tailor useful exercises to suit the language situation. For example, it was noted that the definite article is used for giving specific information, particularly when describing aims, procedures, methods, results and conclusions. A very particular use of the definite article in technical-scientific writing was noted in the following: "We find many references to this in *the* literature (not "in literature"), as it means specific literature referring to the research topic in hand and not literature in general.

(ii) the comparison of L1 and L2

There are many cases in which L1 structures influence L2 language usage. Therefore, some examples of L1 interference were underlined in various categories. Indeed, the tendency of Italian students to form long nominal phrases, following the L1 pattern was pointed out, whereas in technical English language, short concise nominal groups are preferred. For example, 'The design of the 3D virtual environment' would be better expressed as 'The 3D virtual environment design', or 'The BCI's cycle of life should be 'The BCI life-cycle' and 'problems of processing business data' could be more concisely expressed as 'business data processing problems'. The final noun in the group is the principal noun and those which precede it qualify it like adjectives, so even if those nouns are in reality plural, they should be singular in the nominal group, e.g. 'high-speed analyses of clusters' is expressed as 'high-speed cluster analyses'.

Some difficulties were encountered in the use of prepositions in specific contexts, due to L1 interference. For example, 'We reached the conclusion that X is the same **as** (not 'of') Y' and 'Theorem C is dependent **on** (not 'by' or 'from') the multiples of t'.

False friends in the specialized native speaker corpora were also discussed and used in the discourse context. They were compared with the words in the learner (L1) corpora to help explain the confusion regarding their meaning. For instance, 'This hypothesis led to the **actual** ('reale/effettivo') procedure', as opposed to 'This hypothesis led to the current (**'attuale'**) procedure' and 'We **realised** ('ci siamo resi conto che') the model was suitable', compared with 'We created (**'abbiamo realizzato'**) a suitable model'.

(iii) the specific Computer Science language tools

The use of uncountable nouns, such as 'software', 'hardware', 'storage' and

'equipment', is frequent in the Computer Science corpus (Pyne *et al*. 1996). Concepts are often condensed into compound nouns, such as 'screenshot', 'spreadsheet', 'username', 'password', 'toolbox', 'tabletop', 'touch-screen' and 'benchmark'. As previously mentioned, nominal groups form a vast part of the specialized discourse of Computer Science: 'trace clustering', 'data streams', 'batch mode classification', 'text processing', 'Cross-Device Interaction', 'Handwriting Recognition' and 'user parameters'.

Moreover, specific adjectives have often been invented for the discourse genre, in order to reduce a longer concept into short explicit terms. An example of this is 'spaghetti-like process models' (i.e., the process models are tangled and confused like spaghetti and therefore are difficult to understand). Some other examples noted by the students are 'a decision-making task', 'a web-based device', 'user-centered design', 'real-world contexts', 'multi-user notions' and 'a walk-up-and-use table-top'. Some other typical examples of this tendency are the noun 'burstiness' to describe the frequent, intermittent appearance – in bursts – of a word in electronic text analysis, and the verb to 'de-anonymize' data, referring to the data-mining strategy in which anonymous data are cross-referenced with other data sources, in order to re-identify them.

An abundant use of acronyms has also been noted: LDA (Latent Dirichlet Allocation); HDP (Hierarchical Dirichlet Prior) and ReFeX (Recursive Feature eXtraction). This typically technical linguistic 'habit' has obviously been adopted to avoid the frequent repetition of long nominal groups and provide a more concise and rapidly comprehensible linguistic code for the reader.

The specialized Computer Science corpus also displays a series of phrasal verbs which are frequently repeated like 'carry out a task', 'key in a text', 'login in/off/out', 'switch on/off' and 'plug into'. It also encourages the creation of new computer science terms such as 'e-business', 'e-zine', 'e-voting', 'iPod', 'iPad', 'iRead' and 'cyberlink' (Remacha Esteras 2011: 88). The common use of specific prefixes is also evident: 'semi-conductor', 'micro-chip', 'hyper-parameters', 'multi-touch' and 'interaction'. Furthermore, there are cases in which specific words, which in General English normally function as nouns, are commonly used as verbs, for example, 'to impact' (meaning 'to have an impact/effect on something') or 'to bias' (meaning 'to give more importance to one aspect rather than another').

Moreover, the analysis of key Computer Science linguistic tools attests to the intricate relationship between Computer Science and Mathematics languages: the use of multiple sign systems like the Greek alphabet, algebra, formulae and graphical representations (Rudd and Bagnardi 2014). Both languages share common 'axiomatic' discourse: expressions such as 'if P, then Q', 'P implies Q', 'P only if Q' and 'P is a sufficient condition for Q'. The word cluster "if and only if" is frequently used in both languages to obtain a biconditional sentence from two sentences (Rudd and Butts 2007: 191-192).

(c) The training of the students

During the course the student training was performed via the following steps:

- various guidelines concerning recurrent grammar points were introduced and practice exercises were given in class to reinforce the PhD students' linguistic skills;

- in order to discover stylistic features, the students were required to perform specially developed activities, such as highlighting expressions in a text used for giving importance to a particular field of study, e.g. 'Extracting **effective features** for nodes of a given graph is **a key step** for many data mining tasks';

- in order to evaluate the progress made, the students were required to produce models of the various sections of their own papers/theses, which were discussed in groups and with the teacher;

- as a final test the students were asked to produce an abstract for one of the target papers, as the abstract had been deliberately eliminated before analysing this paper in class. The aim of writing the abstract was to evaluate the students' performance in using the afore-mentioned structural guidelines and suitable stylistic features of the specific Computer Science discourse. It also tested their comprehension of the technical concepts discussed in the paper and with the discipline specialist;

- as a final self-monitoring evaluation, the students completed a 'Cloze test', which was the real missing abstract, from which some specific lexical items and syntactic structures had been eliminated. The Cloze test not only helped the students to concentrate on specific points that had caused initial problems, but also to summarize the whole content of the specific target paper.

4. Results

The results are encouraging, as a greater fluency in the writing of rhetorical discourse can be observed. Moreover, the sentences are now shorter, more concise and comprehensible. We note that the students have a better command of modal verbs, and also demonstrate an improved capability in using the article, singular and plural nouns, the position of the adjective, key verbs and linking words.

The quality of the PhD students' written work has improved stylistically thanks to the unique analyses and practice activities of the learning object. The students are able to construct their own models for writing a paper, adapting the linguistic style and rhetorical structures noted to their own specific area of research in Computer Science. They have found the models for the various sections of a paper extremely helpful as a basis for writing their own papers and theses.

Moreover, the students have realized that Computer Science language, like any other scientific language, "meets the needs of scientific argument and theory" (Halliday 1985: 2) i.e. recognizing a problem and defining it, collecting and analysing data, making hypotheses, designing experiments and drawing conclusions. They have also noticed how important it is to read other people's work to help them absorb certain rhetorical discourse patterns and produce their own papers/theses. They have acquired a higher level of self-criticism and evaluation of other colleagues' writing, which is essential to the learning process.

The corpus-based learning object approach adopted has proved to be effective for analysing the discourse genre of native speaker writers to enrich the non-native speaker novice writers' specialized corpus. The authors' knowledge of the PhD students' learner

corpus (L1, Italian) has also helped them to gain a fuller understanding of the students' particular linguistic difficulties and to develop suitable guidelines, aimed at solving very specific problems.

5. Conclusions

The aim of this work was to represent a small specialist language corpus within a learning object in a specific English language teaching situation. Through the analysis of Computer Science discourse in the specific field of 'Data Mining', we have attempted to improve the PhD students' English writing skills when elaborating their papers and theses. The approach adopted was not merely 'statistical-quantitative' (concentrating on single word occurrences), but mainly 'qualitative', focusing on recurrent lexical, syntactic and stylistic features frequently found in the target papers examined. Moreover, the analysis was done in the context of a guided 'discovery learning' situation in the classroom, using specific activities developed by the language specialists and aided where necessary by the discipline specialists.

The students found the activities interesting and particularly stimulating, since they were guided to detect relevant features and patterns themselves in the texts and then build models for their own papers/theses on the basis of these discoveries. Thus, their awareness of key corpus tools was certainly heightened and their ability to capture specialized expressions was enhanced. The focus on grammatical and stylistic difficulties related to the comparison of L1 and L2 was appreciated, since the students were made conscious of the divergences and/or concordances between the two languages, also expanding their vocabulary. Important techniques, such as writing definitions, helped the students to acquire some linguistic mechanisms and scientific habits of mind, in order to develop rhetorical, technical discourse. In fact, as can be clearly seen in the written work produced by the students at the beginning, during the course and in the final test, there has generally been a marked improvement in the use of the specialized discourse.

Since a relatively small group of students was involved, the authors were able to analyse the communicative and pragmatic functions of the corpus and observe the students' interaction with the discourse. Moreover, during the learning process, the students were encouraged to discuss and check the models in small groups, thus fostering a collaborative atmosphere and a critical analysis (peer-reviewing) of other people's work.

As a result of the corpus analysis conducted in this paper, we can conclude that Computer Science discourse is a variety of scientific discourse in general, i.e. it has some features that are common to scientific discourse, but at the same time it presents some unique linguistic patterns that distinguish it from any other type of scientific discourse. This is amply proved by the lexical, syntactic and stylistic features revealed during the analysis of the specialized Computer Science texts.

Regarding the size of the language corpus used for the learning object, in the authors' opinion, analysing small corpora does not create a limitation. On the contrary, it has the advantage of extracting more immediate specific, concise information from the target texts. In larger corpora, the information needs to be dealt with on a wider scale and then divided into more specific subcorpora, which could actually slow down the process of

language learning in a specific situation. In fact, the students themselves recognized that using a small corpus had a number of advantages in helping them to quickly capture specific linguistic features in the specialized discourse. Thus, it can be concluded that the use of a small Computer Science language corpus enhanced the students' ability to assimilate and reproduce specific linguistic features and patterns in Computer Science texts on 'Data Mining'.

An interesting idea for further work would be to ask a different group of PhD students to perform electronic concordancing on the same corpus used in this study, in order to compare the results of the linguistic performance with those obtained through 'manual' concordancing in the classroom.

Acknowledgements

The authors wish to thank Prof. Donato Malerba, Full Professor and Coordinator of the PhD Course in the Department of Computer Science, University of Bari, and Prof. Annalisa Appice, Researcher and Assistant Professor in the Department of Computer Science, University of Bari, for their help in selecting the target papers used for the analysis of the specialized corpus. Their comments and suggestions during the development and application of the learning object and in-class explanations were also much appreciated.

References

Bagnardi A. and L. Rudd 2012. English Practice in Computer Science. A multimedia e-learning object for English language studies in the field of Computer Science. In T. Roselli, A. Andronico, F. Berni, P. Di Bitonto and V. Rossano (eds), *Atti del Convegno, Didamatica 2012*, Informatica per la Didattica, Taranto 14/16 maggio 2012: mondo digitale.aicanet.net/2012-2, Didamatica 2012, sezione: Didattica Multimediale: 1-10.

Beck R.J. 2008. *What are Learning Objects? Learning Objects*. Center for International Education, University of Wisconsin-Milwaukee. Retrieved from https://ww4.uwm.edu/cie/learning_objects.cfm?gid=56.

De Saussure F. [1916] 1986. *Cours de Linguistique Générale.* Bibliothèque Scientifique. Published by Charles Bally, Professeur à l'Université de Genève and Albert Sechehaye, Professeur à l'Université de Genève with the collaboration of Albert Riedlinger Maître au Collège de Genève. Critical edition prepared by Tullio De Mauro. Postface by Louis-Jean Calvet. Payot, Paris 106, Boulevard Saint Germain 1986.

De Saussure F. 2002. *Ecrit de Linguistique Générale*. Paris: Editions Gallimard.

Gavioli L. 2006. *Exploring Corpora for ESP Learning*. Amsterdam / Philadelphia: John Benjamins.

Gilquin G., S. Granger and M. Paquot 2007. Learner corpora: the missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6/4: 319-335.

Glasman-Deal H. 2010. *Science Research Writing: For Non-Native Speakers of English*. London: Imperial College Press.

Gorni G. 1990. *Lettera Nome Numero, L'Ordine delle Cose in Dante*. Bologna: Il Mulino.

Gotti M. 2008. *Investigating Specialized Discourse*. Peter Lang: Bern.

Granger S., J. Hung and S. Petch-Tyson 2002. Computer learner corpora. *Second Language Acquisition and Foreign Language Teaching*. Amsterdam / Philadelphia: John Benjamins.

Halliday M.A.K. 1985. Systemic background. In J.D. Benson and W.S. Greaves (eds), *Systemic Perspectives on Discourse*, vol.1 (ADPS15): 1-15.

Halliday M.A.K. 1988. On the language of physical science. In M. Ghadessy (ed.), *Registers of Written English: Situational Factors and Linguistic Features* (OLS). London: Pinter: 162-178.

Hyland K. 2008. Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18/1: 41-61.

Kendall M., N. Wakefield and R. Delbridge 2007. Enhancing the library and information management curriculum through reusable learning objects. *ITALICS: Innovations in Teaching & Learning in Information & Computer Sciences* 6/2: 52-61.

Malerba D., M. Ceci and M. Berardi 2005. A hybrid strategy for knowledge extraction from biomedical documents. *ICDAR Workshop on "Neural Networks and Learning in Document Analysis and Recognition"*, Seoul, Korea, August 2005.

McEnery T., R. Xiao and Y. Tono 2006. *Corpus-based Language studies - An Advanced Resource Book*. London: Routledge.

McEnery T. and R. Xiao 2010. What corpora can offer in language teaching and learning. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, vol. 2. London/New York: Routledge: 364-380.

Pyne S. and M.T. Allene (eds), M. Ashby (phonetics editor) 1996. *Oxford Dictionary of Computing For Learners of English*. Oxford: Oxford University Press.

Raimes A. 1999. *Keys for Writers - A Brief Handbook* (second edition). Boston, MA: Houghton Mifflin Company.

Remacha Esteras S. 2011. *Infotech, English for Computer Users*. 4th edition. Cambridge: Cambridge University Press.

Rudd L. and A. Bagnardi 2014. English in Computer Science and Mathematics: rhetorical devices and strategic vocabulary, *ALAPP 2014, 4th International Conference of Applied Linguistics and Professional Practice,* Switzerland, Geneva 10-12 September 2014, www.unige.ch/alapp2014.

Rudd L. and M.P. Butts 2007. *English in Computer Science and Mathematics.* 2nd edition. Bari: DigiLabs Publishers.

Sinclair J.M. (ed.) 2004. *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Swales J.M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales J.M. 1994. *Academic Writing for Graduate Students: A Course for Non-native Speakers of English*. Michigan: University of Michigan Press.

Swales J.M 2004. *Research Genres. Explorations and Applications*. Cambridge: Cambridge University Press.

Tarantino M. 2004. Epistemic and dialectic pathway to knowledge, meaning and language advancement. *LSP & Professional Communication* 4/1: 69-88.

Tarantino M. 2005a. Pragmatic and cognitive presuppositions across discourse spheres. *LSP & Professional Communication* 5/2: 73-95.

Tarantino M. 2005b. A holistic approach to language: exo-textual sources of meaning and communication, *ESP Across Cultures* 2: 119-134.

Wallwork A. 2011. *English for Writing Research Papers*. New York: Springer Publishing.

Willis J. 1998. Concordances in the classroom without a computer. In B. Tomlinson (ed.), *Materials Development in Language Teaching*. Cambridge: Cambridge University Press: 30-45.

Windish U. 1985. *Le raisonnement et le parler quotidiens*. Interdisciplinary research group for social reasoning and discourse (Head: U. Windish), University of Geneva. Work and publication funded by the National Swiss Funding for Scientific Research. Lausanne: The Age of Man Editions.

Yoon H. and A. Hinveyla 2004. ESL student attitudes towards corpus use in L2 writing. *Journal of Second Language Writing* 13/4: 257-283.

Zobel J. 2014. *Writing for Computer Science*. 3rd edition, London: Springer Verlag.

APPENDIX

A Co-training Strategy for Multiple View Clustering in Process Mining

Annalisa Appice and Donato Malerba, Member, IEEE

Abstract - Process mining refers to the discovery, conformance and enhancement of process models from event logs currently produced by several information systems (e.g. workflow management systems). By tightly coupling event logs and process models, process mining makes it possible to detect deviations, predict delays, support decision making and recommend process redesigns. Event logs are data sets containing the executions (called traces) of a business process. Several process mining algorithms have been defined to mine event logs and deliver valuable models (e.g. Petri nets) of how logged processes are being executed. However, they often generate spaghetti-like process models, which can be hard to understand. This is caused by the inherent complexity of real-life processes, which tend to be less structured and more flexible than what the stakeholders typically expect. In particular, spaghetti-like process models are discovered when all possible behaviors are shown in a single model as a result of considering the set of traces in the event log all at once. To minimize this problem, trace clustering can be used as a preprocessing step. It splits up an event log into clusters of similar traces, so as to handle variability in the recorded behavior and facilitate process model discovery. In this paper, we investigate a multiple view aware approach to trace clustering, based on a co-training strategy. In an assessment, using benchmark event logs, we show that the presented algorithm is able to discover a clustering pattern of the log, such that related traces result appropriately clustered. We evaluate the significance of the formed clusters using established machine learning and process mining metrics.
Index Terms - Clustering, Co-training, Multiple view learning, Process mining

Published in: *IEEE Transactions on Services Computing*, USA, California, 2015.

Online Active Inference and Learning

Josh Attenberg
Polytechnic Institute of NYU
Brooklyn, NY
josh@cis.poly.edu

Foster Provost
NYU Stern School of Business
New York, NY
fprovost@stern.nyu.edu

Abstract - We present a generalized framework for active inference and the selective acquisition of labels for cases at prediction time, in lieu of the estimated labels of a predictive model. We develop techniques within this framework for classifying in an online setting, for example, for classifying the stream of web pages where online advertisements are being served. Stream applications present novel complications because (i) we don't know at the

time of label acquisition what instances we will see, (ii) instances repeat based on some unknown (and possibly skewed) distribution. To address the complications, we combine ideas from decision theory, cost-sensitive learning, online density estimation, and we introduce a method for on-line estimation of the utility distribution that allows us to manage the budget over the stream. The resulting model indicates which instances to label, so that by the end of each budget period, the budget is best spent (in expectation). We test the method on streams from a real application. The main results show that: (1) our proposed approach to active inference on streams can indeed reduce error costs substantially over alternative approaches, (2) more sophisticated online estimations achieve larger reductions in error. We then discuss the setting of simultaneously conducting active inference and active learning. We argue and provide some support that our expected-utility active inference strategy also selects good examples for learning.

Published in: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2011*: 186-194.

Experiments with Non-parametric Topic Models

Wray Buntine
Monash University
Clayton, VIC, Australia
wray.buntine@monash.edu

Swapnil Mishra
RSISE, The Australian National University
Canberra, ACT, Australia
swapnil.mishra@anu.edu.au

Abstract - In topic modelling, various alternative priors have been developed, for instance, asymmetric and symmetric priors for the document-topic and topic-word matrices, respectively, the hierarchical Dirichlet process prior for the document topic matrix and the hierarchical Pitman-Yor process prior for the topic-word matrix. For information retrieval, language models exhibiting word burstiness are important. Indeed, this burstiness effect has been shown to help topic models as well, and this requires additional word probability vectors for each document. Here we show how to combine these ideas to develop high-performing non-parametric topic models exhibiting burstiness, based on standard Gibbs sampling. Experiments are done to explore the behavior of the models under different conditions and to compare the algorithms with those previously published. The full non-parametric topic models with burstiness are only a small factor slower than standard Gibbs sampling for LDA and require double the memory, making them very competitive. We look at the comparative behaviour of different models and present some experimental insights.

Published in: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* New York (KDD 2014: 881-890).

It's Who You Know: Graph Mining Using Recursive Structural Features

Keith Henderson
Lawrence Livermore Lab
keith@llnl.gov

Brian Gallagher
Lawrence Livermore Lab
bgallagher@llnl.gov

Lei Li & Leman Akoglu
Carnegie Mellon University
{leili,lakoglu}@cs.cmu.edu

Tina Eliassi-Rad
Rutgers University
eliassi@cs.rutgers.edu

Hanghang Tong
IBM Watson
htong@us.ibm.com

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Abstract - Given a graph, how can we extract good features for the nodes? For example,

given two large graphs from the same domain, how can we use information in one to do classification in the other (i.e., perform across-network classification or transfer learning on graphs)? Also, if one of the graphs is anonymized, how can we use information in one to de-anonymize the other? The key step in all such graph mining tasks is to find effective node features. We propose ReFeX (Recursive Feature eXtraction), a novel algorithm, that recursively combines local (node-based) features with neighborhood (egonet-based) features; and outputs regional features – capturing "behavioral" information. We demonstrate how these powerful regional features can be used in within-network and across-network classification and de-anonymization tasks – without relying on homophily, or the availability of class labels. The contributions of our work are as follows: (a) ReFeX is scalable and (b) it is effective, capturing regional ("behavioral") information in large graphs. We report experiments on real graphs from various domains with over 1M edges, where ReFeX outperforms its competitors on typical graph mining tasks like network classification and de-anonymization.

Published in: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2011*: 663-671.