Towards a specialised corpus of organized crime and terrorism*

Pedro Ureña Gómez-Moreno
Ángela Alameda Hernández
Ángel Felices Lago
Universidad de Granada

Abstract

This paper shows the process of compilation of a specialised multilingual corpus for the fields of global organized crime and terrorism. This tailor-made corpus is currently being compiled with a twofold aim. Firstly, it is intended to serve as a repository in the population of a legal subontology within the core ontology included in FunGramKB (Functional Grammar Knowledge Base). Secondly, a more immediate goal of the corpus is to serve as the basis for the creation of a glossary of the abovementioned domains. This terminological glossary will be implemented to show a dictionary interface to the definitions of the specialised terminology therein contained, while it will also be suitable for application in automatic translation, information retrieval and related NLP (Natural Language Processing) tasks. This paper focuses on the methodology followed for compiling the corpus while also showing the process of terminological extraction.

Keywords: corpus linguistics, terminology, FunGramKB, legal ontology.

Resumen

En el presente trabajo mostraremos el proceso de compilación de un corpus especializado de inglés en materia de delincuencia organizada y terrorismo. Este corpus ha sido construido con un doble objetivo. En primer lugar, el de servir como repositorio para la población de una subontología jurídica como parte integrante de la ontología general contenida en la base de conocimiento FunGramKB (Functional Grammar Knowledge Base). En segundo lugar, tendrá como objetivo más inmediato servir como base lingüística para la creación de un glosario especializado dirigido a los dominios legales mencionados. El glosario presenta una interfaz a modo de diccionario y ofrece un acceso inmediato a las definiciones de los términos especializados. Por otro lado, el glosario constituye una herramienta de aplicación en traducción automática, recuperación de información y otras tareas relacionadas con el PLN (Procesamiento de Lenguaje Natural). Este artículo se centrará en los aspectos metodológicos relacionados con la compilación del corpus especializado, así como en el proceso de extracción terminológica.

Palabras clave: lingüística de corpus, terminología, FunGramKB, ontologías legales.

* This paper is part of the research project entitled *Elaboración de una subontología terminológica en un contexto multilingüe (español, inglés e italiano) a partir de la base de conocimiento FunGramKB en el ámbito de la cooperación internacional en materia penal: terrorismo y crimen organizado* funded by the Spanish Ministry of Science and Innovation. Code: FFI2010-15983.

Introduction

The upsurge of new social demands coming from educational, academic and other professional contexts has brought about an increasing necessity for the creation of a more exhaustive body of specialised knowledge as well as the development of versatile tools for accessing such knowledge. In this regard, the area of criminal law is especially in need of systems that can manage technical information effectively and ultimately improve the completion of daily tasks concerning information processing. One of the main tools for the organisation and conceptualisation of legal information lies in the creation of ontologies, customarily defined as a computational apparatus representing a system of concepts which capture a certain domain of knowledge (Gómez-Perez, Fernández-López & Corcho, 2004). For the purpose of populating the ontology, terminologists and knowledge engineers have benefited much from the compilation of corpora. This paper deals with the methodological issues concerning the building of the Global Crime Term Corpus (GCTC), which intends to satisfy some of the demands in the area of legal knowledge and legal ontology-building. This corpus, which is currently being compiled, aims to lay as the basis in the process of creation of a subontology which will ultimately serve the purpose of structuring information for easy retrieval in professional contexts and the solving of problem-oriented tasks in real situations. The corpus plans to be multilingual so as to provide conceptual information that will be a reference in facilitating international communication and reducing confusion in international settings. The chosen languages for this initial stage of our project are English, Spanish and Italian. The present paper deals with the methodological process followed in the creation of this corpus and the initial results obtained in the terminological analysis of the corpus.

This paper is divided into two parts. The first contains a presentation of the main aspects of the methodology followed in the process of collecting a body of relevant texts of the legal sub-domains under concern. We will focus on the procedural aspects involved in the sampling and compilation of a balanced and representative corpus. The second part shows the functioning of the term extracting system included in FunGramKB (Functional Grammar Knowledge Base) Suite (Periñán-Pascual & Arcas-Túnez, 2005), which was used for the retrieval of the units integrating the specialised lexicon. We will also deal with the terminological aspects concerning the mining of n-grams and the creation of lexicological definitions of term candidates.

FunGramKB: An overview

Since the corpus we are currently compiling will help to populate a legal subontology within the core ontology included in FunGramKB, it seems sensible to provide some basic notions about this system. FunGramKB is a multipurpose knowledge-base for natural language processing (NLP) systems. It is multipurpose because it can be used in many different computational tasks, and also because it has been designed to work with any human language. FunGramKB is structured into three main models: a conceptual model, whose main axis component is a nuclear ontology; a lexical model containing all the different lexica of the different languages, and a grammatical model, which comprehends the different grammatical rules that are specific to each language (Mairal

Usón & Periñán-Pascual, 2009; Periñán-Pascual & Arcas-Túnez, 2004; Periñán-Pascual & Mairal Usón, 2009).

Corpus compilation: design and collection

The initial stages in the process of corpus compilation included a number a decisions and selections that would help us to collect and organize the intended corpus coherently and efficiently (Bowker & Pearson, 2002; Koester, 2010).

To begin with, the legal subdomain of organized crime and terrorism was selected both for its international relevance nowadays and for the stated scarce references particularly with the purpose of working with and populating ontologies. This way, the terms extracted from our corpus will populate a specific-domain satellite ontology of the main general ontology in the system of FunGramKB. The corpus was named General Crime Term Corpus (GCTC) because of the legal field it deals with.

A main issue accounted for in the development of a representative corpus relates to the selection of the sources, that is, the entities and documentary repositories whose texts serve to feed the corpus. For the building of the GCTC, two main sources were considered: international institutions and academic works. Hence, in order to collect a significant amount of texts dealing with our chosen topic, a number of institutions were selected from which to obtain such documents. It had to be institutions concerned not only with general legal aspects but particularly related to questions on terrorism and organized crime as international issues. In addition, they had to be entities with official web pages with free access, in order to get the documents in digital format and, hence, facilitate their computer processing. After an intensive survey, the institutions that were finally searched were the UN (United Nations), The Criminal Court of Justice, Europol (European Police Office), Eurojust (European Judicial Cooperation Unit), and OSCE (Organization for Security and Co-operation in Europe), among others. These organizations and legal representatives in the field of fight against organised crime and terrorism offer a rich representation of the technical issues and the specialised vocabulary that is officially used in the issuing of reports and law enforcement in the combat against criminal and terrorist acts in a global and communitarian setting. At the same time, it could be these and similar institutions that would ultimately be benefited from the outcomes of this work and research. Other sources, such as academic reference works and journal articles, were also considered due to the assumed high concentration of specialised terms in their texts.

Next, once the relevant documents were selected and downloaded, a series of hand and semiautomatic editing tasks were required in order to filter out typographical mistakes resulting from the reformatting of original formats (usually pdf) to plain text. These preparatory pre-processing of the texts was necessary because of the characteristics of the term extractor tool, part of the FunGramKB suite, which only works with raw texts. Thus, as a way of example, whenever necessary, manual editing included tasks such as linking words which the automatic text converter had separated or subdividing longer documents when they exceeded file size.

The huge amount of texts collected led the researchers to consider the need to create some sort of organized database where to store all the relevant information related to

the texts. This database would serve as a storage system for easy text identification and access whenever necessary. Each entry includes information about the language of the document, the type of text, the source it had been taken from, a brief description of the text's content and a final identification code. Indeed, to facilitate text identification, we decided to follow a coherent routine for text coding. This way, each text was labelled with a code which included basic information such as language, topic, text type and content. As an example, in the following code "ETRep mass destruction", "E" stands for English, "T" stands for Terrorism, "Rep" stands for Report and "mass destruction" is a brief summary of the text's content.

These initial steps in the process of compilation of the GCTC corpus produced significant and satisfactory results which are described below. Once the English component of the GCTC corpus was completed and closed, the following stage comprised the extraction of specialised terms, whose process is described in section 4.

Characteristics of the GCTC

As has already been mentioned above, in the near future the GCTC will be made up of three different components representing the English, Spanish and Italian languages, these two latter not yet developed. Work so far on the English component presents a corpus which comprises texts on organised crime and terrorism, and is represented by a rich variety of different text types, such as technical reports, declarations, committee decisions or acts. The English component alone contains approximately 6 million words and it is made up of above 600 texts. Two main criteria were taken into account for the compilation stage; namely, balance and representativeness. Balance alludes to the fact that the corpus contains a fair representation among all its constituent parts. In this regard, the English component of the GCTC is formed out of 621 texts out of which the 49% deal with organized crime, 34% deal with terrorism, while the remaining 17% contains texts concerning these two topics simultaneously. Representativeness refers to the fact that a corpus should ideally be a fair sample of the language it pursues to capture. In this regard, the robustness of the GCTC resides mainly in the rich variety of texts that it contains and the fairly large amount of words it holds contributing to capturing legal language. The vast array of text types includes reports, agreements, declarations, regulations, acts, treaties, resolutions and journal articles, among others, adding up to a total of 45 different text types.

Term extraction process

FunGramKB Suite includes a term extractor tool for the assisted retrieval of sets of potentially relevant terms for the fields under study. The extractor applies a series of filters to an input corpus, mainly removal of non-textual characters, numbers and punctuation marks. It is upon this cleaned up text that the statistical extraction process operates. FunGramKB Extractor calculates a tf-idf score for each lexical unit in the corpus. As a result, the terminologist can work on a list of candidate terms ranked according to their semantic weight, so that candidates that appear higher in the list are statistically more relevant specialised terms, while elements that show a tf-idf index below 3 are not

statistically specialised. It is important to notice that the extraction process in FunGramKB is semi-automated and that the ultimate decision of what counts as a specialised term relies on the criterion of the terminologist. Figure 1 below shows the main menu of the extractor containing the principal functions of the tool:

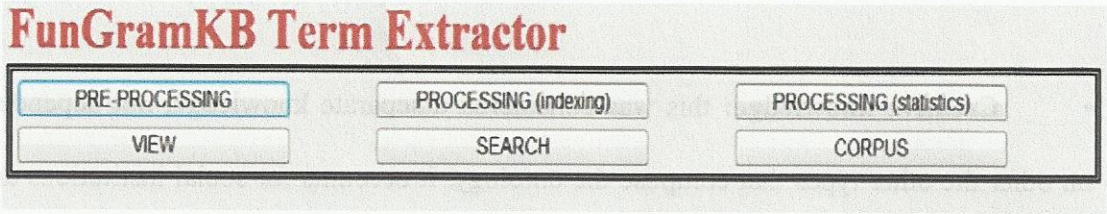

Figure 1. Main Menu of the FunGramKB Extractor

From the top leftmost button: the "Pre-processing" tab contains an area for testing new features for the extractor. The "Processing (indexing)" tab is used for uploading texts of a corpus to the extractor. "Processing (statistics)" is a key function allowing the terminologist to automatically obtain the list of candidate terms from the corpus. "View" allows the terminologist to filter false terms by means of a series of removal options. The "Search" tab is a secondary tool for searching strings of text in a corpus. Finally, "Corpus" shows basic descriptive statistics concerning the number of indexed texts making up a given corpus as well as the number of tokens included. This tab also shows a terminological box containing a list of false candidates that were discarded during the filtering process tackled in the "View" function.

One of the most outstanding features of FunGramKB Extractor lies in its potential for filtering false candidates. The "View" mode contains for each term candidate an option for "simple removal", so that if the terminologist chooses this option an example trigram such as "system term candidate" would be sent to the list of false candidates in "Corpus". More interestingly, the extractor can also make complex removal of lexical bigrams and trigrams. For example, the nested removal of "system term candidate" will result in the removal of "system term candidate" as a trigram as well as in the removal of each component individually ("system", "term" and "candidate") and also the combination of them ("system term", "term candidate").

Preliminary results in the application of FunGramKB emphasize the utility of this approach for term extraction. After uploading to the extractor the English component of the GCTC, which contains roughly five million and a half tokens, and after applying the preparatory filters and the statistical processor, the initial count was reduced to a set of approximately 5,700 candidate terms, that is, a comparatively much smaller quantity of candidate terms. It is important to emphasise that we reached such a reduced set of candidate terms in a short period of time, if compared to other approaches such as manual inspection of concordances or collocations.

Conclusions

This paper has presented the main guidelines towards the compilation of a multilingual corpus for English, Spanish and Italian on criminal and terrorist matters. As explained above, a series of criteria were considered for the collection of a relevant set of texts included in the corpus. In this regard, the texts chosen for the English section were collected from a number of prestigious and renowned international organizations whose documents on criminal issues were appropriate for the purpose of retrieving relevant specialised terms. Another claim of this paper is that FunGramKB Extractor has proved promising results in what concerns the semi-automated retrieval of specialised terms. As has been discussed, FunGramKB Extractor is a tool included in FunGramKB Suite that serves the purpose of assisting the terminologist in a number of different tasks involved in the filtering of irrelevant false candidates, such as boilerplate removal, filtering of common terms and, more importantly, the calculation of semantic weight.

References

— Bowker, L. & J. Pearson (2002). *Working with Specialized language: A practical guide to using corpora*. USA / Canada: Routledge.

— FunGramKB Suite: http://www.fungramkb.com [21/09/2011].

— Gómez-Pérez, A., M. Fernández-López & O. Corcho (2004). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. London / New York: Springer-Verlag.

— Koester, A. (2010). "Building small specialised corpora" in A. O'keeffe & M. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics*, 66-79. US / Canada: Routledge.

— Mairal Usón, R. & C. Periñán-Pascual (2009). "The anatomy of the lexicon component within the framework of a conceptual knowledge base". *Revista Española de Lingüística Aplicada* 22: 217-244.

— Periñán-Pascual, C. & F. Arcas-Túnez (2004). "Meaning postulates in a lexico-conceptual knowledge base", *15th International Workshop on Databases and Expert Systems Applications*, 38-42. Los Alamitos (California).

— Periñán-Pascual, C. & F. Arcas-Túnez (2005). "Microconceptual-Knowledge Spreading in FunGramKB" in *Proceedings on the 9th IASTED International Conference on Artificial Intelligence and Soft Computing*, 239-244. Anaheim-Calgary-Zurich: ACTA Press.

— Periñán-Pascual, C. & R. Mairal Usón (2009). "Bringing Role and Reference Grammar to natural language understanding". *Procesamiento del Lenguaje Natural*, 43: 265-273.