Using Language Resources in Humanities research

Marta Villegas, Nuria Bel, Carlos Gonzalo, Amparo Moreno and Nuria Simelio
Universitat Pompeu Fabra
Universitat Autònoma de Barcelona

E-mail: marta.villegas@upf.edu, nuria.bel@upf.edu, carlos.gonzalo@upf.edu.com, amparo.moreno@uab.cat, Simelio.Sola@uab.cat

Abstract

In this paper we present two real cases, in the fields of discourse analysis of newspapers and communication research which demonstrate the impact of Language Resources (LR) and NLP in the humanities. We describe our collaboration with (i) the *Feminario* research group from the UAB which has been investigating androcentric practices in Spanish general press since the 80s and whose research suggests that Spanish general press has undergone a dehumanization process that excludes women and men and (ii) the "*Municipals'11 online*" project which investigates the Spanish local election campaign in the blogosphere. We will see how NLP tools and LRs make possible the so called 'e-Humanities research' as they provide Humanities with tools to perform intensive and automatic text analyses. Language technologies have evolved a lot and are mature enough to provide useful tools to researchers dealing with large amount of textual data. The language resources that have been developed within the field of NLP have proven to be useful for other disciplines that are unaware of their existence and nevertheless would greatly benefit from them as they provide (i) exhaustiveness -to guarantee that data coverage is wide and representative enough- and (ii) reliable and significant results -to guarantee that the reported results are statistically significant.
The research reported in this paper is part of the activities carried out within the CLARIN (Common Language Resources and Technology Infrastructure).

Keywords: humanities research, NLP web services, massive data

1. Introduction & aims

CLARIN[1] is a large-scale pan-European project which aims at creating a persistent and stable infrastructure addressing the needs of European researchers in the fields of Humanities and Social Sciences. In this framework, one of the main tasks we undertook was the construction of a Spanish national demonstrator that we call CLARIN-es-lab. The eventual demonstrator is a virtual laboratory that collects a set of NLP tools deployed as web services and it is designed for researchers who deal with huge amounts of textual primary data and need to do some kind of text analysis.

One of the goals of our research was to investigate the problems to be addressed when introducing NLP tools and LRs to Humanities and Social Sciences (HSS). Thus, we addressed the *Feminario* research group from the UAB[2] (Women and Mass Culture Feminary, from the Department of Journalism and Communication Sciences) and the "*Municipals'11 online*" project as we understood they were two good real examples from HSS fields in the sense that, they deal with textual primary data, they perform manual text analysis that could be potentially automated and, though their research was qualitatively brilliant and innovative, the amount of data they could analyze and manage was not big enough.

The rest of the paper is organized as follows. In Section 2, we present the two use cases focusing on the user requirements. In Section 3, we present our collaboration as far as data collection is concerned. In Section 4, we describe our contribution to the analysis of data. Finally, in Section 5, we list the conclusions of our experience and summarize the lessons learnt.

2. Cases Study

The *Feminario* research group at the Universitat Autònoma de Barcelona has been investigating androcentric practices in Spanish general press since late 80s. They have developed an innovative methodology to analyze the way mass media represent the social reality. This research provided the basis for various investigations on the representation of women and men in the press since the Spanish transition.

The research concluded that the Spanish press, especially the press considered "serious", has ossified into a male-centered view that expels citizens as active protagonists from public debate. According to the authors, the so called general press does not take into account the whole population. On the contrary, general press reduces its attention and preference to a portion of the population: adult males from dominant classes and people that act on

[1] CLARIN in Spain has received funds from the 7FP of the EU (FP7-INFRASTRUCTURES-2007-1-212230), the Spanish Misterio de Educación y Ciencia (CAC-2007-23) and Ministerio de Ciencia e Innovación (ICTS-2008-11 and ACI2009-0995) and the Generalitat de Catalunya, this project is committed to the integration of the Catalan language in CLARIN by the development of a demonstrator.
[2] Universitat Autònoma de Barcelona (www.uab.es)

the stage of power, and excludes the remaining women and men as if their contributions were insignificant for social functioning.

During this research, the group undertook two main experiments. In the first one, they analyzed the headlines of 15 issues of Sunday magazines of two general Spanish newspapers from 1974 to 1999. In the second one, they analyzed the headlines of 11 issues of four daily newspapers. They manually collected, loaded and analyzed a total of 776 and 3298 headlines respectively. The compilation process lasted for one year and a half whereas the whole project lasted for three years.

As a follow-up of their initial research, they started a new project called "Guía para Humanizar la Información". In this case, the group analyzed a selection of the headlines in the front and back pages of four Spanish newspapers. The set of headlines was manually taken from the issues of the first Friday of February in 1984, 1994 and 2004 respectivelly. The whole project lasted for one year.

The "*Municipals'11 online*" is an e-communication and e-politics project that analyses the elections campaigns in Spain. The main objective of the project is to analyze new tendencies in e-politics and to demonstrate the impact of internet on electoral process. They need to access and analyze texts on the internet daily in order to detect changes and tendencies and to provide a diachronic study.

In both cases the efforts devoted to both the gathering and exploitation of primary data were tremendous. The user requirements can be summarized as follows:

- The need to reduce the time invested in data harvesting and analysis processes.
- The need for really significant amounts of data that provide their experiments with (i) exhaustiveness -to guarantee that data coverage is wide and representative enough- and/or (ii) reliable & significant results -to guarantee that the reported results are statistically significant.
- The need for tools that perform intensive and automatic text analysis processes, even on a daily basis.

3. Allowing for massive data

One of the problems most researchers in HSS face every day is the location and gathering of primary data. In our case, we benefit from the fact that our institution has a diachronic corpus that collects press data daily since 2002 from El País (www.elpais.com). Basically, the crawling process goes as follows: (i) every day at the same time, the system connects to the web page of El País (www.elpais.com) and accesses the "printed version" section[3] (ii) the system locates the URLs of the list of news collected in that section (iii) downloads the news and (iv) identifies the title and paragraphs and removes the html tags.

The first column in Figure 1 shows the total amount of gathered headlines for each year. There we can see that the total amount of headlines included in the new experiment is 151,657 and, more important, this represents the whole universe of data, that is: all new headlines published in the "printed version" of El País from 2005, 2007 and 2010.

Taking into account that in previous manual analyses, the total amount of headlines included in the experiment were 4074, the impact of the increase is obvious. In previous experiments, our researchers selected small portions of randomly chosen "representative" data.

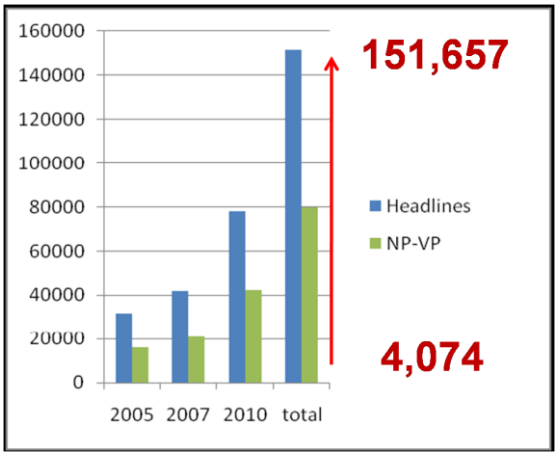

Figure 1: Number of headlines gathered for the current experiment. The First column shows the total amount of headlines. The second column shows the amount of headlines with a NP-VP syntactic pattern

In previous experiments, the lack of digital methods made it impossible to analyze all data and forced our researchers to select small portions of randomly chosen "representative" data. Now we have the whole universe of data.

People from the *Muncipals'11 online Project* need daily data from the Catalan political blogosphere in order to get a diachronic corpus. In this case the task was harder as we had to implement ad hoc html and boilerplate removal. Crawling data from the web and extract plain data may be a laborious task. However the time invested was necessary and fruitful as the researchers were able to perform automatic texts analysis that could not be done otherwise. We eventually managed to automatically access to 459 blogs from the Catalan political blogosphere. This process allowed us to get more than 8000 posts.

[3] http://www.elpais.com/diario. Note that the "Printed Section" is just a section of the digital newspaper not the real printed newspaper.

4. NLP and massive computation

The *Feminario* research group developed a test to evaluate what they call the "informative glance". Essentially the test addressed four basic questions: (i) who focuses the news –this includes the author and/or source, (ii) what women and men are focused –that is, the protagonists of the news (iii) in which actions are they involved and (iv) in what kind of scenarios. The ultimate goal was to determine who the news were talking about and in what terms.

In previous analyses, researchers read the headlines and filled in a form with the questions above. Once all headlines were analyzed they started the statistical analysis. The objective was to help them in the tedious and time consuming task of filling in the forms by providing tools that could automatize the process.

Though the process could not be fully automatized for (ii) and (iii) above CLARIN NLP tools proved really useful. Essentially, the tasks that CLARIN tools addressed were:

- the identification of both the protagonists and actions of the headlines
- the classification of protagonists (both syntactically and semantically) and
- the analysis of associations between protagonists and their corresponding actions in order to find significant associations.

The sequence of tasks performed by CLARIN tools can be summarized as follows:

1. Automated dependency syntactic analysis of the headlines.
2. Selection of the headlines with NP-VP structure.
3. Identification of the subject as the protagonist of the action
4. Identification of the main verb as the action of the headline
5. Semantic classification of subjects (that is, protagonists). This includes common nouns as well as proper names.

1 to 4 was automatically done in just a few minutes using Freeling service[4]. The deep-syntactic parser was able to analyse nearly 100% of the headlines.

As we can see in Figure 1 above, around the 50% of the headlines in each year follow the sentence structure NP-VP. The rest of the headlines were essentially noun phrases and adjectival phrases (among other less frequent structures). Notice that our analysis only considered those headlines with NP-VP structure as we were looking for "subjects" (protagonists) and "verbs" (actions). This means that a total amount of 79.611 headlines were eventually included in the experiment.

The semantic classification of proper names was also performed using Freeling tools. In Freeling, Named Entity Classification (NEC) distinguishes between four types: *locative*, *organization*, *person* and *miscellaneous*. As fully reported in Carreras et al. (2002), the Spanish Freeling NEC system has a precision of 81.38% and a recall of 81.40% overall types. More important in our case, Freeling NEC gets precision of 84.71% and a recall of 93.47% when considering the person type alone.

The semantic classification of subject common nouns was automatically performed via lexicon look up. This lexical look-up process was automated and it took less than two minutes. The process was able to semantically classify more than 91% of the protagonists.

Figure 2 summarizes this classification. The first two columns distinguish between proper nouns vs. common nouns. The second two columns distinguish between human vs. non-human protagonists[5]. The figures demonstrate that as far as protagonists are concerned (i) proper nouns are more frequent than common nouns (60% and 40% respectively) and (ii) non-human protagonists are always more frequent than human protagonists (55% and 45% respectively).

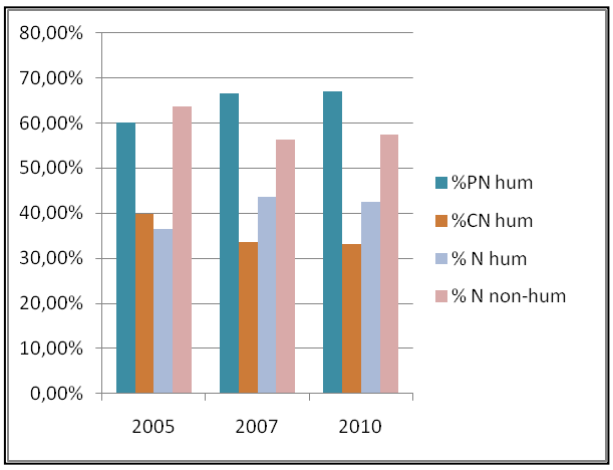

Figure 2: Classification of protagonists

Once we had the protagonists and actions identified and classified, we provided different quantitative analysis tools. These include basic quantitative measures such as different word counts, measures of vocabulary richness, sample representativeness, vocabulary distribution and different measures of association. We were able to identify most statistically significant associations between subject and verbs, compare occurrences between masculine and feminine words and word forms, identify

[4] Freeling is an open source language analysis tool suite developed by the TALP research center in the Universitat Politècnica de Catalunya (http://nlp.lsi.upc.edu/freeling/). This tool is deployed as a web service in the virtual laboratory at http://clarin-es-lab.org.

[5] This includes both, proper nouns and common nouns.

most frequent actions associated to most frequent protagonists, etc.

People from the *Municipals'11* online project performed daily manual analysis of the Catalan political blogosphere. They were particularly interested in key word identification and quantitative analyses.

We decided to use TfIdf (term frequency–inverse document frequency) in order to identify most significant words for each blogosphere in contrast to the others[6]. The experiment demonstrated we got better results if we applied some sort of term identification before TfIdf calculus. In this way the TfIdf considered multiword terms identified in previous task as single tokens. The workflow we designed was as follows:

1. tokenization of input text,
2. term identification using Ted Pedersen's ngrams package[7] and
3. TfIdf calculus.

Once the workflow was ready, the researchers could execute it daily. Our users really appreciated the possibility to run successive experiments on new data. Note that they needed to perform a daily analysis of the blogosphere in order to identify tendencies and changes.

The hardest problem we addressed had to do with text handling tasks and the fact of dealing with user generated content. We eventually managed to automatically access to 459 blogs from the Catalan political blogosphere. This process allowed us to get more than 8000 posts.

All the NPL tools used in these experiments were deployed as (SOAP) web services using the Soaplab2 software [8] . They are part of the virtual laboratory developed at the UPF and can be found at http://clarin-cat-lab.org and http://clarin-es-lab.org. These virtual laboratories give access to the Soaplab Web Client (a Graphic User Interface [GUI] called Spinet) from where services can be executed. In addition, UPF web services can be combined into workflows to perform complex tasks using a workflow editor like Taverna[9]. Taverna is an open source Workflow Management System used to design and execute scientific workflows. Taverna comes with a Soaplab plug-in and allows easy access and execution of services.

5. Results and discussion

It is still difficult to find and access primary resources. This is especially important taking into account that data are increasing dramatically in a distributed manner. Humanities' researchers need to have easy access to primary data and must have enough processing power to perform the required operations and analysis.

We could collaborate with researchers from UAB and run this experiment as we had a corpus of Spanish press since 2002. Note, however, that the corpus only includes one masthead and only covers a period of 10 years. CLARIN NLP services prove efficient when processing large corpora but large corpora are not always available.

When crawling data from the blogosphere for the *Municipals'11 online*" project we faced additional problems: html cleaning and boilerplate removal is still a time consuming task. Unfortunately, the use of semantic web and micro formats is not extended enough and we had to develop 'ad hoc' cleaning processes. In addition, most of the blogs in the experiment were open to user comments. This means we faced the problem of dealing with user generated content.

Language technologies have evolved a lot and are mature enough to provide useful tools to researchers dealing with large amount of textual data. The language resources that have been developed within the field of Natural Language Processing have proven to be useful for other disciplines that are unaware of their existence and nevertheless would greatly benefit from them. A common effort has to be made to create the necessary synergies to allow researchers from other disciplines to use NLP tools. A mutual understanding of the real needs of these researchers and the applications available to fulfil them is still needed. NLP tools allow automating processes that were manually carried out in the past.

NLP tools have interesting results, help us in boring tasks and allow automating analyses but they work on plain textual data. When applying NLP tools to massive data the real bottleneck is not only getting the desired data but getting them clean and ready to be processed. Most NLP tools require plain text input data in a given character encoding. However, real data tends to be formatted data and often come from several sources with different formats and encodings. Getting plain text out of 'real' data is very time consuming. Users of NLP tools may be disappointed if they need to address such things.

6. Acknowledgements

This work was developed in the framework of the CLARIN project. CLARIN was supported by the Spanish Ministerio de Ciencia y Tecnología (ACI2009-0995), the Departament d'Innovació, Universitats i Empresa of the Generalitat de Catalunya, and the European Union (FP7-INFRASTRUCTURES-2007-1-212230).

7. References

Xavier Carreras and Lluís Màrquez and Lluís Padró. (2002) Named Entity Extraction using AdaBoost. In

[6] Catalan political blogosphere was split into four partitions according to the political party involved.
[7] http://ngram.sourceforge.net/
[8] http://soaplab.sourceforge.net/soaplab2/
[9] http://www.taverna.org.uk/

Proceedings of CoNLL Shared Task, pg. 167--170. Taipei, Taiwan.

D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services.," *Nucleic Acids Research*, vol. 34, iss. Web Server issue, pp. 729-732, 2006.

Amparo Moreno. (2007). ¿De quién hablan las noticias? Guía para humanizar la información. Icaria, Colección Akademia. Barcelona. ISBN: 978-84-7426-955-0.

Amparo Moreno. (2007). De qué hablamos cuando hablamos del hombre. Treinta años de crítica y alternativas al pensamiento androcéntrico. Icaria, Colección Akademeia. Barcelona. Páginas: 375. ISBN: 978-84-7426-956-3. Depósito legal: B-49.136-2007

Amparo Moreno. (1998). La Mirada informativa. Barcelona, Bosch Casa Editorial

Lluís Padró and Miquel Collado and Samuel Reese and Marina Lloberes and Irene Castellón. (2010). FreeLing 2.1: Five Years of Open-Source Language Processing Tools *Proceedings of 7th Language Resources and Evaluation Conference* (LREC 2010), ELRA. La Valletta, Malta. May, http://nlp.lsi.upc.edu/freeling.

Senger M., Rice P., Bleasby A., Oinn T., Uludag M.. (2008). "Soaplab2: more reliable Sesame door to bioinformatics programs", In *The 9th annual Bioinformatics Open Source Conference*.

IJ Taylor, E Deelman, DB Gannon, M Shields (Eds) (2006). *Workflows for e-Science: Scientific Workflows for Grids*. Springer-Verlag New York, Inc. Secaucus, NJ, USA.