Computational Linguistics in Costa Rica: an overview*

Jorge Antonio Leoni de León
Escuela de Filología, Lingüística y Literatura
Universidad de Costa Rica
Ciudad Universitaria Rodrigo Facio, San Pedro Montes de Oca
San José, Costa Rica
antonio.leoni@ucr.ac.cr

Abstract

This paper aims to bring a general overview on the situation of Computational Linguistics in Costa Rica, particularly in the academic world.

1 Introduction

Costa Rica is a Central American country, well known for the abolition of its army in 1949 and for its policies in favor of the conservation of ecologically important areas. Investments in Education brought a good development of several higher educational institutions, most of them public and open to new students through an academic selection system. As a consequence, one of the main products of exportation is software (outsourcing, for example). The growing importance of Computational Linguistics (CL) in the last years has made developers to start using CL technologies at work, but unfortunately, often, they are on their own for the lack of academic structures supporting development. In the Table 1 and in the Figure 1 we provide comparative data on *Information and Communication Technologies* (ICT) for Costa Rica, Brazil, Mexico and the United States both taken from the World Economic Report for 2010 (Dutta and Mia, 2010), in order to help the reader to get a better picture of the context on which CL is growing up in Costa Rica.

This paper aims to bring a general account of CL in Costa Rica. This article is divided into three parts. The section 2 starts with some early CL articles in Costa Rica before covering academic and research infrastructure in the country. The section 4 briefly acknowledges the presence of CL industries in the country. Finally, the section 5 presents some of the current tendencies in research and teaching. The claims in this paper are not exhaustive and they only represent the Author's opinion, who aims to be as objective as possible, but who does not have a complete knowledge on the structures, institutions and companies involved in computational issues in Costa Rica. Therefore, all the topics are raised from personal interviews and experiences .

2 Academic infrastructure

Since the 90's, CL raised interest in main academic institutions in Costa Rica, particularly in the field of Artificial Intelligence, but this interest was not continued and well structured in time. Some researchers moved to other academic areas in computation or even to industrial research. So, in the *Instituto Tecnológico de Costa Rica*[1] (TEC) we found the first citations related to CL, mainly on number recognition (Helo and Sell, 1995), knowledge representation (Araya, 1992), connectionism (Vargas, 1991) and unification grammars (Vargas, 1992), these last from a philosophical perspective.

In Costa Rica, we found more than 40 universities and research institutes. But, computational research in the country is principally done in the TEC and in the University of Costa Rica.[2] At this moment, CL

*Thanks to Sharid Loáiciga, Natalia Bermúdez, Prof. Gabriela Barrantes, Prof. Hugo Mora Poltronieri and Prof. Álvaro de la Osa for their suggestions and comments. All the gaps and mistakes in this paper are entirely mine.

[1]Web site: http://www.tec.cr/. Visited: 03/28/2010.

[2]These institutions are public universities; in Costa Rica the best standards in higher education are found in the public insti-

Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pages 40–45, Los Angeles, California, June 2010.

Variable	Costa Rica	Brazil	Mexico	United States
Market Environment				
Venture capital availability	2.68	2.73	2.39	4.17
Availability of latest technologies	4.66	5.29	4.58	6.58
State of cluster development	3.58	4.25	3.76	5.45
Political and Regulatory Environment				
Laws relating to ICT	4.06	4.43	3.90	5.54
Intellectual property protection	3.54	3.04	3.19	5.44
Infrastructure Environment				
Secure Internet servers (hard data)	98.75	23.67	15.67	1173.66
Electricity production (hard data)	1997.70	2259.80	2380.84	14309.62
Availability of scientists and engineers	4.74	4.24	3.64	5.60
Quality of scientific research institutions	4.63	4.22	3.71	6.18
Tertiary education enrollment (hard data)	25.34	29.99	26.93	81.68
Education expenditure (hard data)	4.06	4.44	5.47	4.79
Accessibility of digital content	4.56	4.85	4.53	6.33
Internet bandwidth (hard data)	8.55	20.83	2.81	111.22
Individual Readiness				
Quality of math and science education	4.34	2.71	2.58	4.47
Quality of the educational system	4.69	3.01	2.80	4.85
Business Readiness				
Company spending on R&D	3.75	3.79	2.90	5.63
University-industry collaboration in R&D	4.25	4.06	3.48	5.90
Government Readiness				
Government prioritization of ICT	4.93	4.44	4.25	5.62
Government procurement of advanced technology products	4.00	3.68	3.28	4.77
Importance of ICT to government vision of the future	4.44	4.15	3.98	4.91
Individual Usage				
Personal computers (hard data)	23.10	16.12	14.10	78.67
Business Usage				
Capacity for innovation	3.45	3.90	2.78	5.49
Government Usage				
High-tech exports (hard data)	25.88	5.79	12.25	19.84
Government success in ICT promotion	4.37	4.40	3.83	5.19
ICT use and government efficiency	4.49	4.64	4.37	5.26
Presence of ICT in government agencies	3.88	5.06	4.44	5.81
World rank 2009–2010 (over 133 economies)	49	61	73	5

Table 1: *Comparative data on Information and Communication Technologies (ICT) (Dutta and Mia, 2010)*

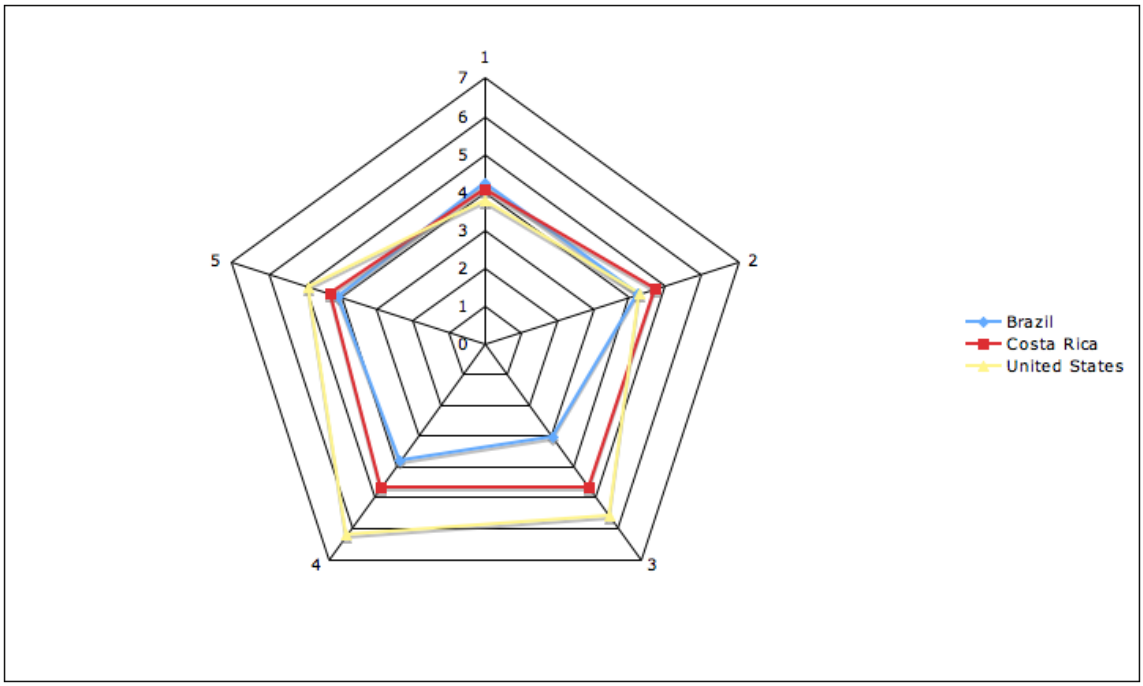

Figure 1: *Benchmarking on chosen ICT data for Costa Rica, Brazil and the United States (Dutta and Mia, 2010)*

is mainly found in the University of Costa Rica as part of the academic structure. This institution, as far as we are concerned, has two structures covering research: schools and research institutions. The former are mainly oriented towards teaching; the latter are exclusively devoted to research. CL teaching is carried out in the *School of Philology, Linguistics and Literature*[3] and in its associated postgraduate program on Linguistics[4]. Research is concentrated in the Instituto de Investigaciones Lingüísticas (INIL)[5] at the Facultad de Letras[6]. INIL is under the administrative supervision of the *Vicerrectoría de Investigación*[7], which approves project financing. All the scientific responsability falls on the members and commissions of the research institutions. At this moment, CL is attached to INIL's program ELEXHICÓS. Where a two year project is under development, with the part time assistance of a researcher. This project aims to create the bases for a hybrid parser.

In the Instituto de Investigaciones en Ingeniería (INII), we found a program on cognition and language[8]. In spite of the name, their research addresses mostly behavioral issues.

There exists the possibility of getting funding from governmental institutions, like CONARE[9], but they are meager. Nevertheless, it seems there are good options within inter-university collaboration, but this would require a long term coordination and planification.

In Costa Rica, research funding can be improved if real applications are proposed, for example educational and developmental tools. Other areas of interest are indigenous and minority languages, like Bribri and Chinese. In this sense, there are no computational projects officially inscribed in INIL at this moment, but ideas to start projects on this areas are waiting for the next inauguration of the Natural Language Processing Laboratory in the University of Costa Rica, so there are no papers to cite for the moment. Nevertheless, any inquiry on indigenous lan-

tutions.

[3]Escuela de Filología, Lingüística y Literatura.

[4]Web site: http://www.linguistica.ucr.ac.cr. Visited: 03/28/2010.

[5]Site: http://inil.ucr.ac.cr/. Visited: 02/15/2010.

[6]Roughly "Faculty of Humanities".

[7]In English, the University's Bureau for Research.

[8]Site: http://iniiserver.inii.ucr.ac.cr/picl/. Visited: 02/15/2010.

[9]Site: http://www.conare.ac.cr/. Visited: 02/15/2010.

guages can be addressed to Prof. Carla Jara Murillo, director of the Linguistic Department.

3 Growing needs, growing interest

In the 90's the project *Estudios de Lexicografía Hispano-Costarricense*[10] (ELEXHICÓS)[11] at the *School of Philology, Linguistics and Literature* of the University of Costa Rica was born as an initiative to do lexicographical research on contemporary Costarican Spanish and to publish dictionaries based on modern scientific methods and oriented towards different publics and usages.

Lexicographical studies appeal to large corpora in order to document accurate word values and guarantee their usefulness. The need of computational resources was felt from the beginning of ELEXHICÓS. The Murillo and Sánchez (1993) on lexical and syntactic maturity (language acquisition) is a good example of a research where this need of language computational tools is present. However, statistical dictionaries (Morales, 2009) can only be done by electronic means because of the huge calculations involved. For a long time, ELEXHICÓS counted on the limited, but important, support of the Centro de Informática (Center of Informatics) of the University of Costa Rica, but the need of developing its own resources and technologies imposed itself.

In 2002 the *School of Philology, Linguistics and Literature* of the University of Costa Rica opened the course *Tecnología y Producción Textual*[12], where text processing technologies like LaTeX, XML, HTML, CSS and Perl are taught for applications in the Humanities field. Some interesting proof-of-concept projects have been proposed as part of the course activities (Arroyo Molina, 2009; Enciso Bahler, 2008; Fuentes Vargas, 2008). As an initiative of ELEXHICÓS and the *School of Philology, Linguistics and Literature* in order to develop CL, the University of Costa Rica approved a grant for a Ph.D. on Computational Linguistics, which was fulfilled at the *Laboratoire d'Analyse et de Technologie du Langage*[13] of the University of Geneva. As part of this initiative a full time Professorship on CL was opened for 2010 and since 2009 CL is taught as postgraduate facultative course for the Master on Linguistics. Other courses on Formal Linguistics and Natural Language Processing (NLP) were accepted and will be part of the offer in 2010. Additionally, the *School of Philology, Linguistics and Literature* approved the creation of a NLP laboratory (project number 021-A9-734 of the *Vicerrectoría de Investigación*, main resarcher, Prof. Jorge Antonio Leoni de León, assistant researcher, Prof. Carla Jara Murillo), which is expected for 2010, and should support research and teaching, especially because a course on data processing for undergraduate students on Linguistics and Philology is under consideration.

Computational graduate students have showed their interest on CL courses at the *Postgraduate Program in Linguistics*[14]. Some of them come to the course looking for knowledge on CL, since because of their jobs they need a good understanding of NLP technologies. Although we lack of details about the job they do with NLP, the works of Berrocal Rojas (2009) and Cedeño Baltodano (2009) for the postgraduate program on Computational Sciences illustrate the growing interest in the field. At the moment, only one student started, this year, her Master's thesis on CL at the Postgraduate Program on Linguistics at the University of Costa Rica.

Off universities' campus, it's important to mention the *Centro Nacional de Alta Tecnología*[15] (CENAT), which has projects sharing similarities with CL, but with totally different aims. For example, we can cite the projects on human memory modelisation (in collaboration with the *Programa de Investigaciones en Fundamentos de la Educación a Distancia*, PROIFED, UNED) and an adaptation of computational learning methods to a parallel and distributed processing platform. CENAT has contacts with several organizations at international level. CENAT is a state inter-university research institution on super computation.

[10] In Spanish *Studies on Hispanic-Costarican Lexicography*.

[11] Site: http://www.lexicografia.ucr.ac.cr. Visited: 02/15/2010.

[12] Technology and text production.

[13] Language Technology Laboratory (Site: http://www.latl.unige.ch). Visited: 02/10/2010.

[14] Site: http://www.linguistica.ucr.ac.cr/. Visited: 02/15/2010.

[15] National Center of High Technology.

4 Commercial application of Language Technologies

We are aware of postgraduate students working in related areas in another countries, but we do not have details about their plans for the future. Nevertheless, national industry has not waited to have graduated specialists in order to start the commercialization of NLP related software. For example, Wordmagic Software[16] developed symbolic machine translation software Spanish–English and Tecapro presented an orthographic correction software[17]. Because of secrecy in the industry, it is difficult to know how many companies appeal to NLP technologies. Enterprise are also make themselves present in the field by scientific grants, in 2007 Juan Rodríguez won a prize proposing a glove allowing translation between sign language and Natural Langue.[18]

5 Research interests and collaboration issues

Presently, the idea is that Computational Linguistics would play a role, at different levels, in Lexicography and Spanish as a second language (L2). This leads to the creation of parsing systems and large corpora, where collaboration is desired. Another interesting area for CL is dialectology, where the tradition, as in Lexicography, of large collaborative initiatives exists. The new NLP laboratory will have workstations exclusively dedicated to research and the creation and storage of large copora as an intensive international initiative could be possible.

The signing of specific collaborative agreements between academic institutions is the preferred way to accomplish international research projects. This facilitates the approval of fundings.

6 Conclusions

In Costa Rica, CL is finding its path. At this moment it is mainly an academic interest, but, as we saw, there already are industrial products, where, eventualy, graduated students could find professional employment.

In the University of Costa Rica, where CL is rapidly growing up, collaborative initiatives are possible in a specifc and well defined frame. Additionally, many students are coming to CL, with the new academic offer. And this tendecy can increase in the next years.

In this moment, there is no CL community in Costa Rica. Before we are able to build it, we need to create a solid ground where a CL community could firmly stand up. This is starting to happen at the University of Costa Rica, where individual initiatives begin to gather around the recent course on CL taught by Prof. Jorge Antonio Leoni de León. We hope this movement will continue with the next opening of the NLP Laboratory at the *School of Philology, Linguistics and Literature* in the University of Costa Rica. This laboratory will fill the lack of equipment to make research on CL. Nevertheless, the need of a permanent research group will be there. This is an understandable situation in the sense that CL is a very new branch in the University and especially in *liberal arts*. This allows us to think that students will incorporate to CL studies and projects once the equipment and the offer on courses will be normal. It is expected that this year (2010), *School of Philology, Linguistics and Literature* will have a full time professor on CL, who will dedicate at least $\frac{1}{4}$ of his time to research.

Acknowledgments

Thanks to *Vicerrectoría de Investigación* (http://vra.ucr.ac.cr/vra.nsf), the *Instituto de Investigaciones Lingüísticas* (INIL) and the *Escuela de Filología, Lingüística y Literatura*, for their continued support in the structured research and teaching on Computational Linguistics.

References

C. Araya 1992. “Representación de conocimiento con una lógica modal”. Tiempo Compartido 3(5):21-24.

Amparo Morales. 1986. Léxico básico del español de Puerto Rico. Academia Puertorriqueña de la Lengua Española.

[16] Site: http://www.wordmagicsoft.com. Visited: 02/15/2010.

[17] Sites: http://www.tecapro.com/ContentTeQuita.html and http://www.tecapro.com/TecApro_Historia.pdf. Visited: 02/15/2010.

[18] Read in http://www.intel.com/CostaRica/prensa/Mayo23_07.htm and http://www.nacion.com/ln_ee/2007/mayo/18/ultima-sr1101870.html. Visited: 03/28/2010.

Allan Berrocal Rojas. 2009. Automatización parcial de la revisión de aspectos de precisión, no-ambigüedad y verificabilidad en requerimientos de software escritos en lenguaje natural. Tesis de Maestría. Programa de postrado en Computación e Informática, Universidad de Costa Rica.

Allan Cedeño Baltodano 2009. Comparación del rendimiento de las aplicaciones Toscanaj y Concept Explorer para la construcción de retículas de conceptos. Tesis de Maestría. Programa de postrado en Computación e Informática, Universidad de Costa Rica.

Constanza Enciso Bahler. 2008. "Diccionario Maya–Español" Escuela de Filología, Lingüística y Literatura. Trabajo presentado en el curso FL–1036 Tecnología y Producción Textual. Universidad de Costa Rica.

J. Helo and C. Sell 1995. "Reconocimiento de dígitos mediante redes neurales y lógica difusa". Tecnología en Marcha 12, número especial sobre lógica difusa.

Marielos Murillo Rojas and Víctor Manuel Sánchez Corrales. 1993. Campos semáticos y disponibilidad léxica en preescolares. Revista de Educación. Number 2. Volume 17. Pages 15–25.

Orietta Fuentes Vargas. 2008. Propuesta de formato electrónico del Diccionario Biográfico de Escritores Costarricenses. Trabajo presentado en el curso FL–1036 Tecnología y Producción Textual". Escuela de Filología, Lingüística y Literatura. Universidad de Costa Rica.

Sergio Arroyo Molina. 2008. "Diccionario de Topónimos". Trabajo presentado en el curso FL–1036 Tecnología y Producción Textual". Escuela de Filología, Lingüística y Literatura. Universidad de Costa Rica.

Soumitra Dutta and Irene Mia (Eds.). 2010. *The Global Information Technology Report 2009–2010: ICT for Sustainability*. INSEAD / World Economic Forum. Retrieved March 27, 2010 from http://www.networkedreadiness.com/gitr/main/fullreport/index.html.

Celso, Vargas 1991. "Conexionismo: una alternativa en inteligencia artificial". Mundo de la computación. Vol. 5, No. 28.

Celso, Vargas 1992. "La utilización de formalismos basados en unificación para el análisis de las lenguas naturales". Revista de filología y lingüística de la Universidad de Costa Rica. Vol.18, No.2., p. 71-83.