Computational Linguistics in Brazil: An Overview

Thiago A. S. Pardo[1], Caroline V. Gasperin[1], Helena M. Caseli[2],
Maria das Graças V. Nunes[1]

Núcleo Interinstitucional de Lingüística Computacional (NILC)

[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400 - Centro
P.O.Box 668. 13560-970 - São Carlos/SP, Brazil

[2] Departamento de Computação, Universidade Federal de São Carlos
Rod. Washington Luís, Km 235
P.O.Box 676. 13565-905 - São Carlos/SP, Brazil

{taspardo,cgasperin}@icmc.usp.br, helenacaseli@dc.ufscar.br,
gracan@icmc.usp.br

Abstract

In this paper we give an overview of Computational Linguistics / Natural Language Processing in Brazil, describing the general research scenario, the main research groups, existing events and journals, and the perceived challenges, among other relevant information. We also identify opportunities for collaboration.

1 Brazilian Research Scenario

Computational Linguistics (CL) / Natural Language Processing (NLP) is an emerging and growing area in Brazil. Although there is no consensus, it is traditionally understood as a research field within Artificial Intelligence, gathering researchers mainly from Computer Science/Engineering and Linguistics. There is also modest interaction with Information Sciences area.

In general the CL/NLP area in Brazil started with researchers that finished their PhD abroad and, after coming back, initiated the first CL/NLP projects. Since then, but mainly more recently, the area has experienced some internationalization due to the fact that the number of undergraduate and graduate students that undergo internships on renowned foreign NLP research centers has increased. In Brazil, PhD students have the possibility to take their complete PhD course abroad or, alternatively, only a part of it. In both cases, students may count on Brazilian funding agencies.

The area is more strongly represented and promoted by Brazilian Computer Society (SBC)[1], particularly by its Special Interest Group on NLP (CEPLN)[2], created in 2007. It is interesting that most researchers in Brazil (independent from their background area) do not differentiate CL from NLP, using both terms interchangeably.

Research in Brazil is carried out mainly at public universities and at a few private universities and business companies. Differently from most countries, in Brazil public universities are generally considered the top ones, although exceptions do exist.

Currently, there are no undergraduate courses on CL/NLP in Brazil, therefore researchers in this field come mainly from Computer Science and Linguistics courses. However, there are a few graduate courses on CL/NLP, with both computing and language emphases, such as the MSc and PhD programs at USP/São Carlos[3], UFSCar[4], UNESP/Araraquara[5], PUC-RS[6], and UFRGS[7], among others.

[1] http://www.sbc.org.br
[2] http://www.sbc.org.br/cepln
[3] http://www.icmc.usp.br/~posgrad/computacao.html
[4] http://ppgcc.dc.ufscar.br and http://www.ppgl.ufscar.br
[5] http://www.fclar.unesp.br/poslinpor
[6] http://www.pucrs.br/inf/pos
[7] http://www.ufrgs.br

Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, pages 1–7, Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

Funding for research comes mainly from governmental agencies. Nowadays Brazil has 4 agencies that significantly support research in the country (in this order): CNPq[8] (National Council for Scientific and Technological Development), FAPESP[9] (São Paulo Research Foundation), CAPES[10] (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*), and FINEP[11] (Research and Projects Financing). Private funding is still modest, which reflects the limited interaction between universities and companies. Some of the above agencies have tried to change this scenario by providing special joint university-industry funding programs. For instance, FAPESP and Microsoft Research recently formed a partnership to fund socially relevant projects in the state of São Paulo, e.g., the PorSimples[12] text simplification project. FAPESP also funds special university-company programs, where the research to be developed must be of interest to a company, which, in turn, has to support the research and work together with the researchers.

NLP research in Brazil is varied and deals not only with Portuguese processing, but also with English and Spanish mainly. Given that Portuguese is among the most spoken languages in the world (it is estimated that almost 250 million people speak some variant of Portuguese in the world[13]), research interests on Portuguese processing is shared with other countries, mainly Portugal. In this sense, Portugal has launched an initiative to create and maintain a unified information storage center that indexes resources and publications for/on Portuguese processing. The initiative is the Linguateca project[14], which was officially created in 2002, but initial works date back to 1998. Santos (2009) presents and evaluates the work carried out by Linguateca.

Brazil and Portugal have a history of partnership on Portuguese processing, which formally started in 1993 with the first PROPOR conference (PROPOR event series is introduced in Section 4). We maintain this partnership active by having collaborative projects and promoting joint events.

As far as we know, other Portuguese speaking countries do not have a tradition of CL/NLP research. However, curiously, there are researchers from other non-Portuguese speaking countries that develop relevant research on Portuguese language. For example, to the best of our knowledge, currently the best syntactical parsers for Portuguese were developed by researchers from Denmark and the USA. These researchers actively work with the Brazilian research community.

In what follows, we briefly present the Brazilian research profile (Section 2), the main research groups (Section 3), and the Brazilian events and journals (Section 4). We also report the main challenges for research in Brazil (Section 5) and the collaboration opportunities with other American researchers that we envision (Section 6).

2 Research Profile

In 2009 CEPLN proposed a survey of the status of CL/NLP research in Brazil and published the results during the 7th Brazilian Symposium in Information and Human Language Technology (Pardo et al., 2009). The survey aimed at gathering information both about researchers (such as their location, education level, number of students, etc.) and their research (main research topics, number of funded projects, main challenges, etc.).

The survey was carried out mainly on-line. A call for participation was sent to all known e-mail lists from scientific associations from varied areas. Data was also obtained from the Registry of Latin American Researchers in Natural Language Processing and Computational Linguistics[15].

148 researchers responded to the survey: 35% of these were academic staff with a PhD degree, 16% academic staff with a Master's degree, 1% academic staff with a Bachelors degree, 9% PhD students, 26% Master's students, 14% undergraduate students, and 5% others. Table 1 summarizes the main results of the survey, showing the percentage of answers for each issue. One may see that CL/NLP research is mainly carried out in the south and southeast regions of Brazil.

[8] http://www.cnpq.br

[9] http://www.fapesp.br

[10] http://www.capes.gov.br

[11] http://www.finep.gov.br

[12] http://caravelas.icmc.usp.br

[13] Besides Brazil and Portugal, Portuguese is an official language in Angola, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau, Macau, Mozambique, and São Tomé and Príncipe.

[14] http://www.linguateca.pt

[15] http://ww.d.umn.edu/~tpederse/registry/registry.cgi

Table 1. CEPLN survey

Issues	**Results**
Geographic distribution	48% São Paulo state 18% Rio Grande do Sul state 8% Paraná state 7% Rio de Janeiro state 19% Other states
National collaboration	52% Yes, 48% No
International Collaboration	25% Yes, 75% No
Background area	62% Computer Science 29% Linguistics 9% Other
Supervision of postgraduate students	28% Yes, 72% No
Funded projects	28% Yes, 72% No
Source of funding	43% Federal government agencies 25% São Paulo state government agency 31% Other state government agencies

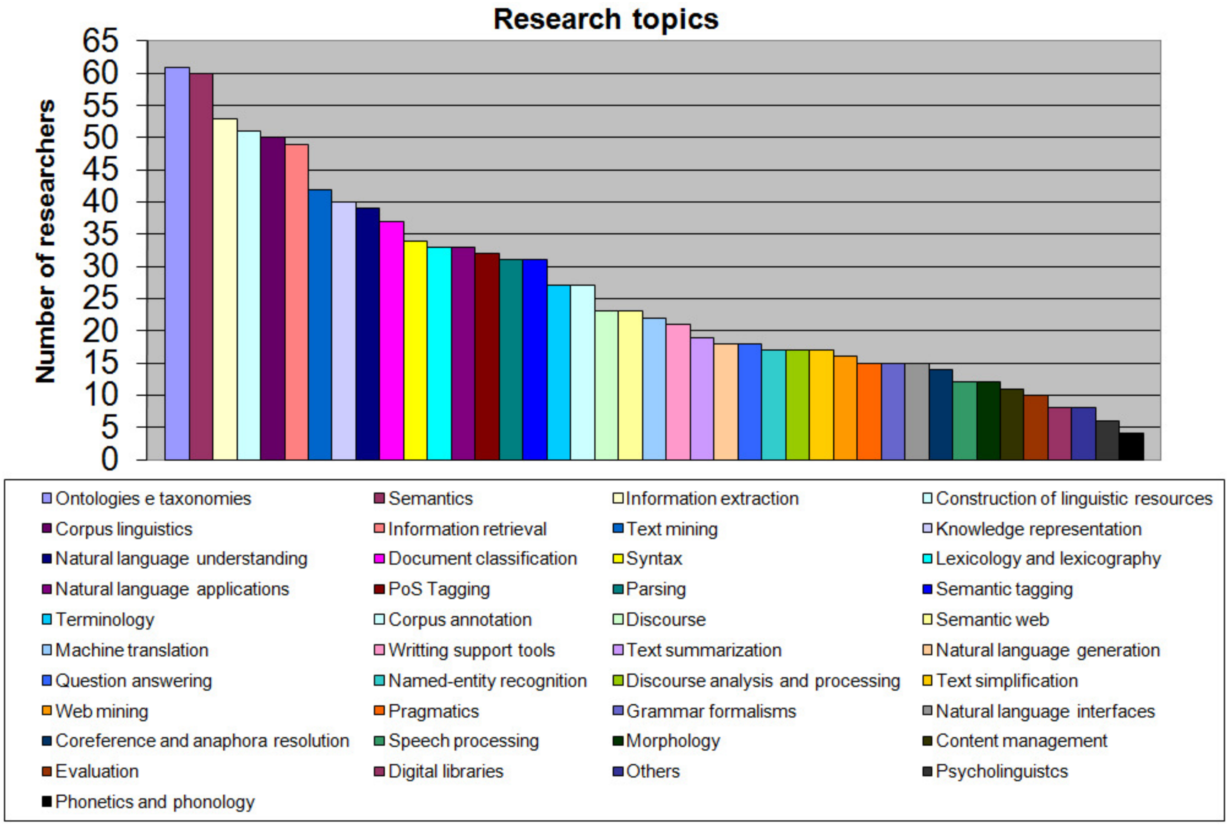

Figure 1. Research topics

The survey also inquired the participants about their research topics. Figure 1 shows the distribution of topics among researchers who responded to the survey. Researchers could mark as many research topics as they wanted. Some topics subsume others, so these were marked more often by respondents.

Ontologies and semantics were the topics marked by most respondents. We believe that there is indeed a significant number of researchers working on them, but we also believe that they are not the main topic of research of most people who listed them. For example, the statistics for "Ontologies" probably also include researchers who simply make use of ontologies in their work and not necessarily develop ontologies or ontology generation methods. Other researchers believe that we are in a changing period, moving from syntax-centered research to semantics-centered research, due to the fact that more recently the community has produced more robust semantic tools and resources, e.g., the first versions of Portuguese language wordnets, as TeP 2.0[16], Wordnet.PT[17], and MWN.PT[18], as well as named entities recognizers, e.g., REMBRANDT[19].

Interestingly, corpus linguistics is one of the hottest topics but, at the same time, it is not seen as a genuine CL/NLP topic: most researchers that indicated corpus linguistics as a research topic marked it as "other area of interest". Some researchers have advocated that CL/NLP area and corpus linguistics should be considered a unique area, while others argue that these areas have different purposes and, therefore, different scientific methods, what would avoid such unification. Text mining is another curious case: research on this theme is mostly carried out by non-CL/NLP researchers, but instead by researchers on general AI and database areas

Based on the publications on the last Brazilian scientific events and on the fact that we personally know most of the CL/NLP researchers in Brazil, we dare to indicate the following topics as the most recurrent ones (in no particular order): text summarization, machine translation, text simplification, automatic discourse analysis, coreference and anaphora resolution, information retrieval, text mining, terminology/lexicon research, ontologies and semantic tagging, and corpus linguistics.

Based on the survey, we estimate that Brazil has about 250 researchers (including students) with interest in CL/NLP area. Although only 148 researchers attended the CEPLN survey, we computed other researchers in the Registry of Latin American Researchers in Natural Language Processing and Computational Linguistics and in the CEPLN e-mail list that did not attend the survey. In general, we estimate that about 35-40 of these are active researchers, whose main topic of research is CL/NLP, and who supervise undergraduate and graduate students on the subject. We also estimate that there are 5-10 researchers on speech processing that actively collaborate with the CL/NLP community.

3 Main research groups

The largest CL/NLP research group in Brazil is NILC (Interinstitutional Center for Research and Development in Computational Linguistics)[20], which includes researchers mainly from University of São Paulo (USP; Computer Science and Physics departments), Federal University of São Carlos (UFSCar; Computer Science and Linguistics departments) and State University of São Paulo (UNESP; Linguistics department). The group was created at 1993.

NILC has a long history of research in CL/NLP, which has thrived since the ReGra[21] project, in which the grammar checker for Portuguese that is currently used within Microsoft Word since its 2000 version was built. In fact, ReGra project was born from a university-industry collaboration, one of the few successful ones in CL/NLP area in Brazil. At the moment most of the research at NILC is concentrated on the following topics: automatic summarization, text simplification, coreference resolution, and terminology. NILC has hosted STIL 2009 (STIL event series is introduced in the next section). NILC also currently holds the presidency of CEPLN.

The NLP group at the Computer Science department at the Catholic University of Rio Grande do Sul (PUC-RS)[22] also has a tradition of research on CL/NLP. Their current projects focus on information retrieval, ontology engineering and anaphora resolution. The group also has research on multi-agent systems applied to NLP tasks and, more recently, on text categorization. The group hosts PROPOR 2010 (PROPOR event series is also introduced in the next section). The group has held the presidency of CEPLN from its creation (2007) until 2009.

The above research group and NILC form the main CL/NLP research vein in Brazil. They have joint research projects and have strong

[16] http://www.nilc.icmc.usp.br/tep2/index.htm

[17] http://www.clul.ul.pt/clg/wordnetpt

[18] http://mwnpt.di.fc.ul.pt

[19] http://xldb.di.fc.ul.pt/Rembrandt

[20] http://www.nilc.icmc.usp.br

[21] http://www.nilc.icmc.usp.br/nilc/projects/regra.htm

[22] http://www.inf.pucrs.br/~linatural

collaboration, constantly hosting graduate students from each other in internship research periods.

There are also other very relevant NLP groups in Brazil that regularly carry out projects on the area. We may cite the Catholic University of Rio de Janeiro (PUC-Rio)[23], Federal University of Rio Grande do Sul (UFRGS), State University of Campinas (UNICAMP), University of the Sinos River Valley (Unisinos), and State University of Maringá (UEM), among others.

4 Events and Journals

The Brazilian Symposium on Information and Human Language Technology (STIL) is the main event on CL/NLP in South America and is in its seventh edition. It is promoted by CEPLN and is carried out since 1993. It is intended to be a forum for gathering everyone with interest in CL/NLP. It happens regularly (every one or two years) and accepts contributions in Portuguese, Spanish and English. Details about the event are available at www.nilc.icmc.usp.br/til.

The International Conference on Computational Processing of Portuguese Language (PROPOR) is an international conference jointly promoted by Brazil and Portugal and is in its ninth edition. It is the main conference with focus on Portuguese language, giving equal space to research on text and speech processing. It is carried out in Brazil and in Portugal interchangeably (every two or three years) and accepts submissions in English only. PROPOR's proceedings are published as part of Springer Lecture Notes series. Details about the event are available at www.nilc.icmc.usp.br/cgpropor.

STIL and PROPOR are the most relevant conferences for researchers in CL/NLP in Brazil. Their last editions received support from NAACL.

AI events are also recurrent forums for CL/NLP researchers. The Brazilian AI events are the Brazilian Symposium on Artificial Intelligence (SBIA)[24] and the National Meeting on Artificial Intelligence (ENIA)[25], also promoted by SBC. They are already in their twentieth and seventh editions, respectively.

Other related events in Brazil are the Corpus Linguistics Meeting (ELC)[26] and Brazilian School on Computational Linguistics (EBRALC)[27], which are in their eighth and third editions, respectively. These events are mainly organized by the Linguistics research community. EBRALC is mainly intended for new students in the area and has been held together with ELC.

Brazilian researchers count mainly on the following journals for national periodical publications:

- JBCS[28] (Journal of the Brazilian Computer Society), which is published by SBC and covers all Computer Science areas, including CL/NLP;
- RITA[29] (Journal of Theoretical and Applied Computing), also of general scope.

It is important to cite Linguamática[30], which is an European initiative to publish CL/NLP research on the Iberian languages.

CEPLN is also organizing a joint journal with other SBC AI-related special interest groups.

5 Challenges

At STIL 2009, the research community discussed challenging issues (raised by respondents of the CEPLN survey) that hamper research on CL/PLN in Brazil. The main issues raised were:

- Lack of large and robust language resources for Portuguese;
- Lack of formal models for linguistic description and analysis of Portuguese;
- Difficulty in attracting students and researchers to the area;
- Lack of multidisciplinary collaboration;
- CL/NLP marginalization in both Computer Science and Linguistics.
- Poor interaction between universities and industry;
- Insufficient funding.

Here we discuss some of these points. Although Portuguese has got state of the art tools (as POS taggers and syntactic parsers) and comprehensive corpora of contemporary written language, there is

[23] www.letras.puc-rio.br/Clic/ogrupo.htm

[24] http://www.jointconference.fei.edu.br/

[25] http://csbc2009.inf.ufrgs.br/

[26] http://www.corpuslg.org/elc/Inicial.html

[27] http://www.corpuslg.org/ebralc/Inicial.html

[28] http://www.springer.com/computer+science/journal/13173

[29] http://www.seer.ufrgs.br/index.php/rita

[30] http://linguamatica.pt

still a need for resources for particular applications or domains. Many researchers feel that Portuguese syntactic parsers (which are considered basic NLP tools) and wordnet-like resources are still too limited, not attending their demands. Brazil also lacks representative spoken corpora, what may be explained by the fact that, in Brazil, written and spoken language processing communities have modest interaction. While written language processing research is reported at SBC events, spoken language processing is mainly conducted under SBrT (Brazilian Telecommunications Society)[31]. PROPOR series have tried to bring together these two communities, fostering joint research and mutual awareness of both research lines.

The lack of formal models for Portuguese linguistic description and analysis was mainly perceived by linguists that work with CL/NLP. In fact, they acknowledge that Brazil has no tradition in carrying out events on these themes, what would eventually harm CL/NLP research. This goes along with Spärck Jones (2007) opinion paper. One first step towards overcoming this lack of formal models for Portuguese description was the Workshop on Portuguese Description[32], carried out together with the last edition of STIL.

Another point that deserves attention is the sentiment that CL/NLP research suffers from marginalization in both Computer Science and Linguistics areas, as it is usually the case for multidisciplinary subjects. We believe this might be fueled by the way research is assessed in Brazil. In Brazil, the quality of research is mainly assessed by the publications generated from it, and publication vehicles from Linguistics are usually rated worse in Computer Science, and vice versa. It is expected that different areas may have different scientific methods and perspectives, as well as it is natural that such differences are mirrored in any evaluation instrument. However, such factors lead some researchers to feel uncomfortable with the multidisciplinary nature of CL/NLP field and the way they are recognized in their own major areas. Many researchers (not only from Brazil, but also from Portugal) have supported that CL/NLP should become a new “major” area, instead of being part of Computer Science or Linguistics.

Concerning insufficient funding, we believe that the main complaints came from Brazilian regions other than south and southeast, which currently concentrate CL/NLP research. In fact, during a lengthy discussion at STIL 2009 about the raised challenges, this issue was dismissed by many participants as non-representative. We believe that the funding situation in each region of Brazil contributes to the status of research on all topics, not particularly CL/NLP, in these regions. While in most Brazilian states researchers have to compete for funding from national agencies, some states (mainly in the southeast region) can rely on strong state-based funding agencies, such as FAPESP, in the state of São Paulo.

6 Opportunities for Collaboration

We believe that there are many opportunities for collaboration on CL/NLP with other researchers in the Americas, mainly due to the fact that the research community in Brazil works not only with Portuguese, but also with English and Spanish.

One first step towards collaboration in Latin America was given in the event CHARLA 2008 (Grand Challenges in Computer Science Research in Latin America Workshop). Organized by several scientific societies (including SBC), the event aimed at contributing to the definition of a long-term research agenda in Latin America with the potential to significantly advance science and motivate the networking of abilities and competencies in Latin America. One of the recognized challenges was “multilinguism”, which involves several CL/NLP topics. CHARLA immediate impact in Brazil was the adaptation of Brazilian CL/NLP events to receive contributions in Spanish, which has a vast number of speakers in Latin America. Contributions in English were already traditionally considered in Brazilian events.

We believe that another important source of collaboration comes from awareness of the ongoing research projects in the Americas. Workshops such as this seem to be a channel for the exchange of information. We envision that initial collaborations may arise within machine translation projects, which naturally already deal with the representative languages of the Americas.

Letting aside technical collaboration, we believe there is room for higher-level concrete actions that

[31] http://www.sbrt.org.br

[32] http://www.ppgl.ufscar.br/jdp/index.html

could foster collaboration in the Americas. These are actions that may increase the visibility of the research done in Latin America, as well as motivate new research. One first action that we envisage is the opening of evaluation challenges and shared tasks to the languages of the Americas other than English. For instance, contests/conferences such as TAC[33], Senseval/SemEval[34], and TREC[35], among others, might make Portuguese/Spanish datasets available, as CLEF[36] has done in its last editions. This has certainly an organizational cost, but it may turn out to be a valuable investment.

Another action that could stimulate the progress of CL/NLP research in Latin America consists of including the proceedings of other American CL/NLP conferences in the ACL Anthology[37], for example, the proceedings of STIL and PROPOR, to mention the Brazilian examples. This could be restricted to conferences that received ACL/NAACL endorsement and/or sponsorship.

While the first action we proposed would make it feasible for more countries to participate in the evaluation contests, the second action would allow the works carried out in these countries to be better known.

In a different strategy, we imagine that it must be possible for regional scientific associations to establish formal partnerships, granting some advantages to associated researchers from the corresponding countries, such as: registration discounts in the CL/NLP conferences from the countries (for instance, ACL/NAACL members would have discounts for registering in Brazilian events, as well as SBC members for ACL/NAACL events); and distribution of relevant publications for members of the associations (for instance, SBC traditionally distributes to its members the JBCS journal, which is considered a prestigious international publication).

Our last idea would be to create a fund (possibly through the associations' partnership) for funding visits for knowledge transfer (1-2 weeks) for researchers and mainly students. These could be an opportunity for studying/working with researchers from other countries that work on topics of interest, as well as for renowned researchers to visit research groups in order to stimulate work on a particular topic. Such opportunities would be very positive for Brazilian students.

We believe that the actions suggested above can lead to a more integrated research scenario in the Americas.

Acknowledgments

The authors are grateful to SBC, CEPLN, FAPESP, and CAPES for supporting this work and the realization of STIL 2009, where part of the data shown in this paper was presented.

References

Pardo, T.A.S.; Caseli, H.M.; Nunes, M.G.V. (2009). Mapeamento da Comunidade Brasileira de Processamento de Línguas Naturais. In the *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology* - STIL, pp. 1-21. September 8-10, São Carlos/SP, Brazil.

Santos, D. (2009). Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, N. 1, pp. 25-58.

Spärck Jones, K. (2007). Computational Linguistics: What About the Linguistics? *Computational Linguistics*, Last Words Section, Vol. 33, N. 3, pp. 437-441.

[33] http://www.nist.gov/tac

[34] http://www.senseval.org

[35] http://trec.nist.gov

[36] http://www.clef-campaign.org

[37] http://aclweb.org/anthology-new