Mapeamento da Comunidade Brasileira de Processamento de Línguas Naturais

Thiago Alexandre Salgueiro Pardo[1], Helena de Medeiros Caseli[2], Maria das Graças Volpe Nunes[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC),
Universidade de São Paulo (USP)
Av. Trabalhador São-carlense, 400 - Centro
Caixa Postal: 668 - CEP: 13560-970 - São Carlos/SP, Brasil

[2]Departamento de Computação (DC)
Universidade Federal de São Carlos (UFSCar)
Rod. Washington Luís, Km 235
Caixa Postal 676 - CEP: 13565-905 - São Carlos/SP, Brasil

taspardo@icmc.usp.br, helenacaseli@dc.ufscar.br, gracan@icmc.usp.br

Resumo. *Relatam-se, neste documento, os resultados do mapeamento da comunidade brasileira de Processamento de Línguas Naturais, realizado entre Maio e Julho de 2009. O mapeamento, realizado pela Comissão Especial de Processamento de Linguagem Natural da Sociedade Brasileira de Computação, foi idealizado com o objetivo de se conhecer melhor a área e, desta forma, permitir o estabelecimento de ações direcionadas para que a área se desenvolva e seja representada apropriadamente no Brasil.*

1. Introdução

No período de 14 de Maio a 10 de Julho de 2009 (aproximadamente 2 meses), a Comissão Especial de Processamento de Linguagem Natural (CEPLN) da Sociedade Brasileira de Computação (SBC) realizou um mapeamento da área no Brasil via uma enquete on-line.

O objetivo era fazer uma "radiografia" da área no Brasil, mais especificamente, quanto à concentração de temas de trabalho, distribuição de pesquisadores, dificuldades enfrentadas na área, número de projetos financiados, e existência de colaboração entre pesquisadores, entre várias outras informações. Com base nessas informações, pretende-se identificar ações direcionadas para que a área se desenvolva e seja representada apropriadamente no Brasil.

Pretendia-se que o público alvo do mapeamento abrangesse pesquisadores de todas as áreas do conhecimento relacionadas direta ou indiretamente com Processamento de Línguas Naturais (PLN). A chamada para participação do mapeamento foi divulgada amplamente em listas de e-mails das áreas de Computação, Lingüística e Ciência da Informação, tanto genéricas quanto específicas da área de PLN. Foram realizadas duas chamadas gerais para participação, com intervalo de aproximadamente um mês entre cada uma.

No total, 148 pessoas responderam à enquete no período citado. Não houve casos de dúvidas ou problemas com o preenchimento da mesma. Algumas pessoas responderam mais de uma vez a enquete; algumas responderam parcialmente. Isso pode acarretar algumas pequenas discrepâncias nos dados (por exemplo, no número total de respostas para alguns itens), que, acreditamos, não são significativas no cenário geral.

Os resultados compilados são apresentados na seção seguinte. Algumas considerações finais são feitas na Seção 3. A enquete completa é reproduzida no Apêndice A.

2. Resultados

A Figura 1 mostra a distribuição de pesquisadores que responderam à enquete em função de suas atividades. Pode-se notar que há muitos doutores e alunos de graduação e de mestrado.

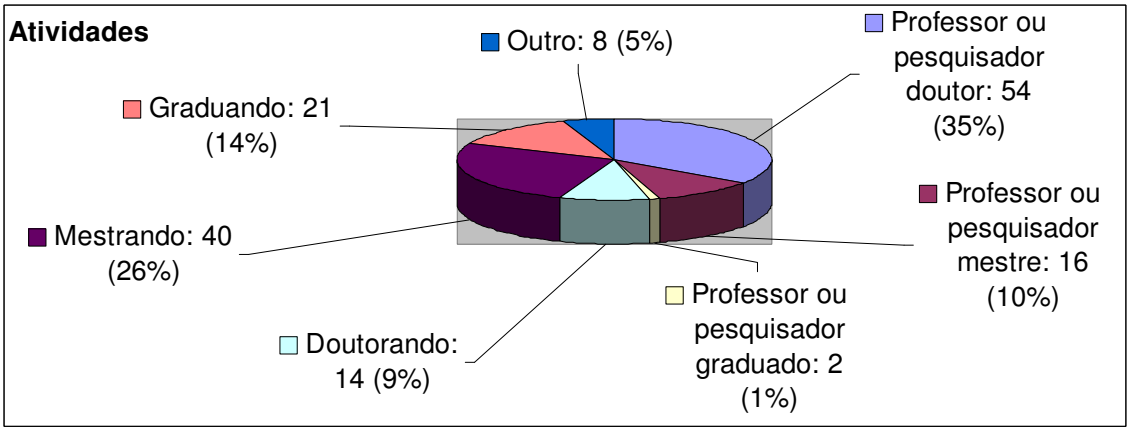

Figura 1. Atividades dos pesquisadores que responderam à enquete

A Figura 2 mostra a quantidade absoluta de pesquisadores por temas de pesquisa. Os temas que predominam são: semântica, ontologias e taxonomias, criação de recursos lingüístico-computacionais, recuperação e extração de informação, e lingüística de córpus. Outros temas de destaque foram interpretação de línguas naturais, representação e modelagem do conhecimento e mineração de textos. A relação de temas foi obtida a partir das chamadas de artigos nos principais eventos e revistas da área. É importante notar que os temas não são disjuntos e que os pesquisadores podem ter tido dificuldade em fazer sua escolha entre os temas oferecidos. Uma mesma pessoa poderia indicar mais de um tema de pesquisa. É interessante verificar que temas de natureza semântica predominam.

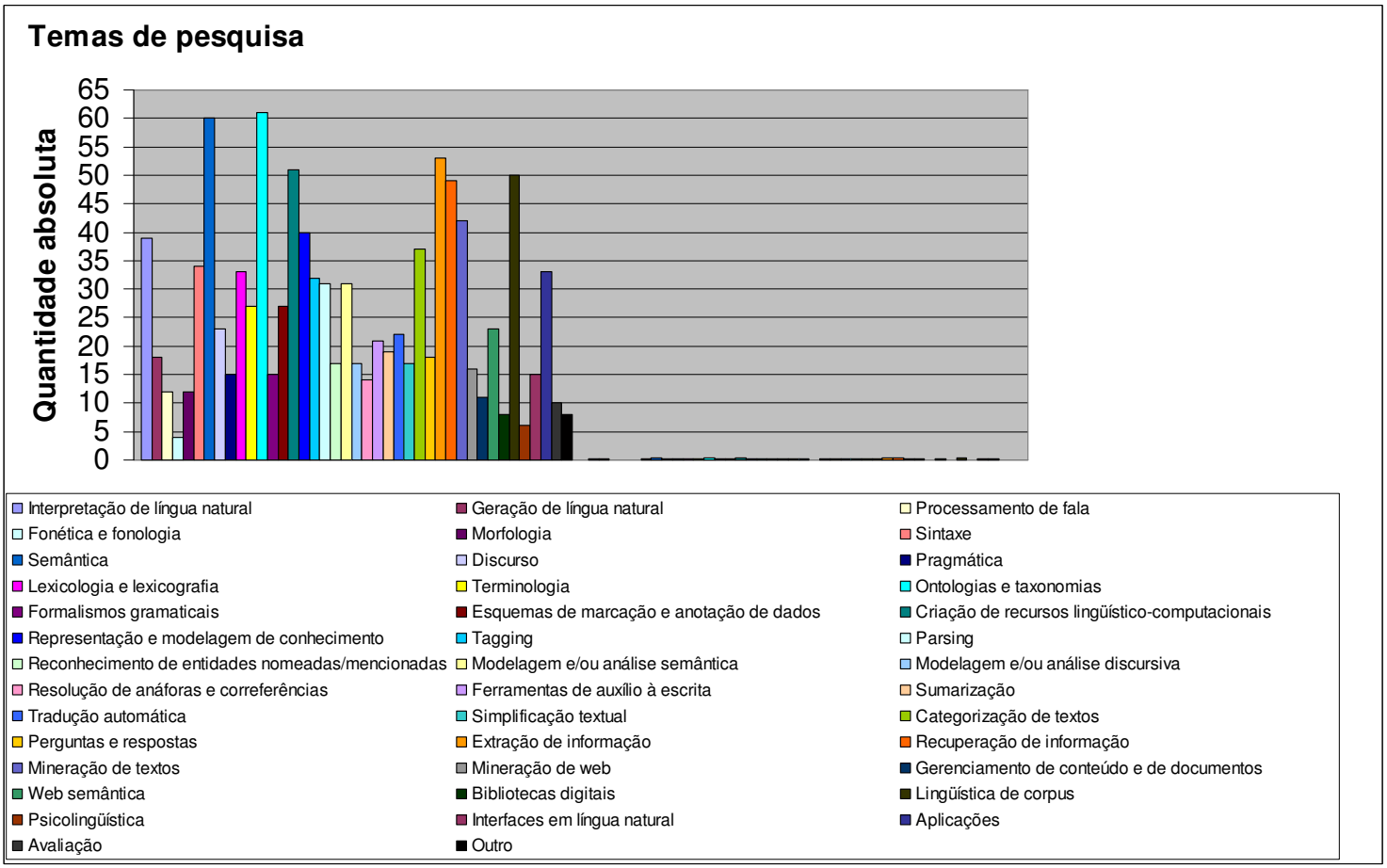

Figura 2. Pesquisadores e temas de pesquisa

A Figura 3 mostra a proporção de pesquisadores que consideram PLN como sua principal área de pesquisa. A Figura 4 mostra a proporção de pesquisadores que, além de PLN, também pesquisam em outras áreas. A Figura 5, por sua vez, exibe a quantidade de pesquisadores que atuam principalmente em outras áreas de pesquisa. Pode-se ver que outras áreas indicadas são IA e Mineração de Dados, também tendo destaque as áreas de Engenharia de Software, Lingüística, Lingüística de Córpus e Tradução.

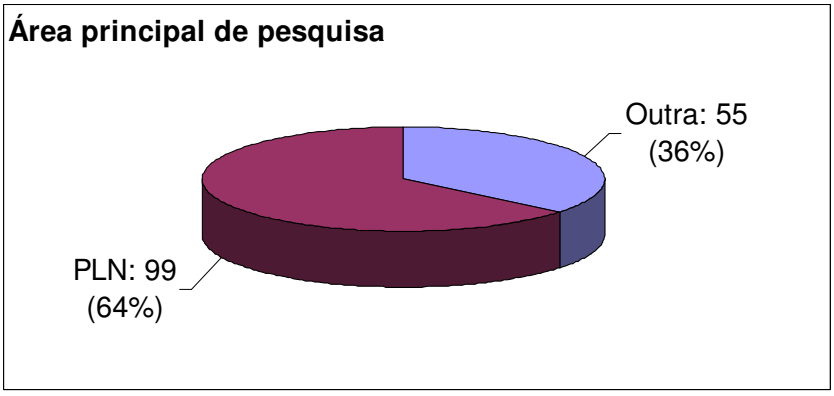

Figura 3. Proporção de pesquisadores cuja principal área é PLN

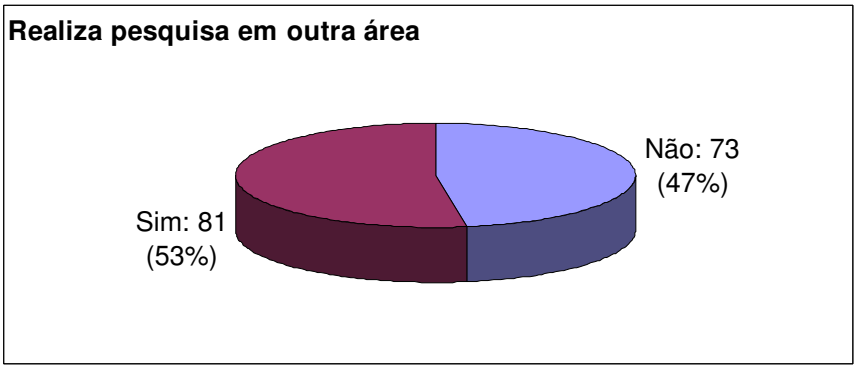

Figura 4. Proporção de pesquisadores que também pesquisam em outras áreas

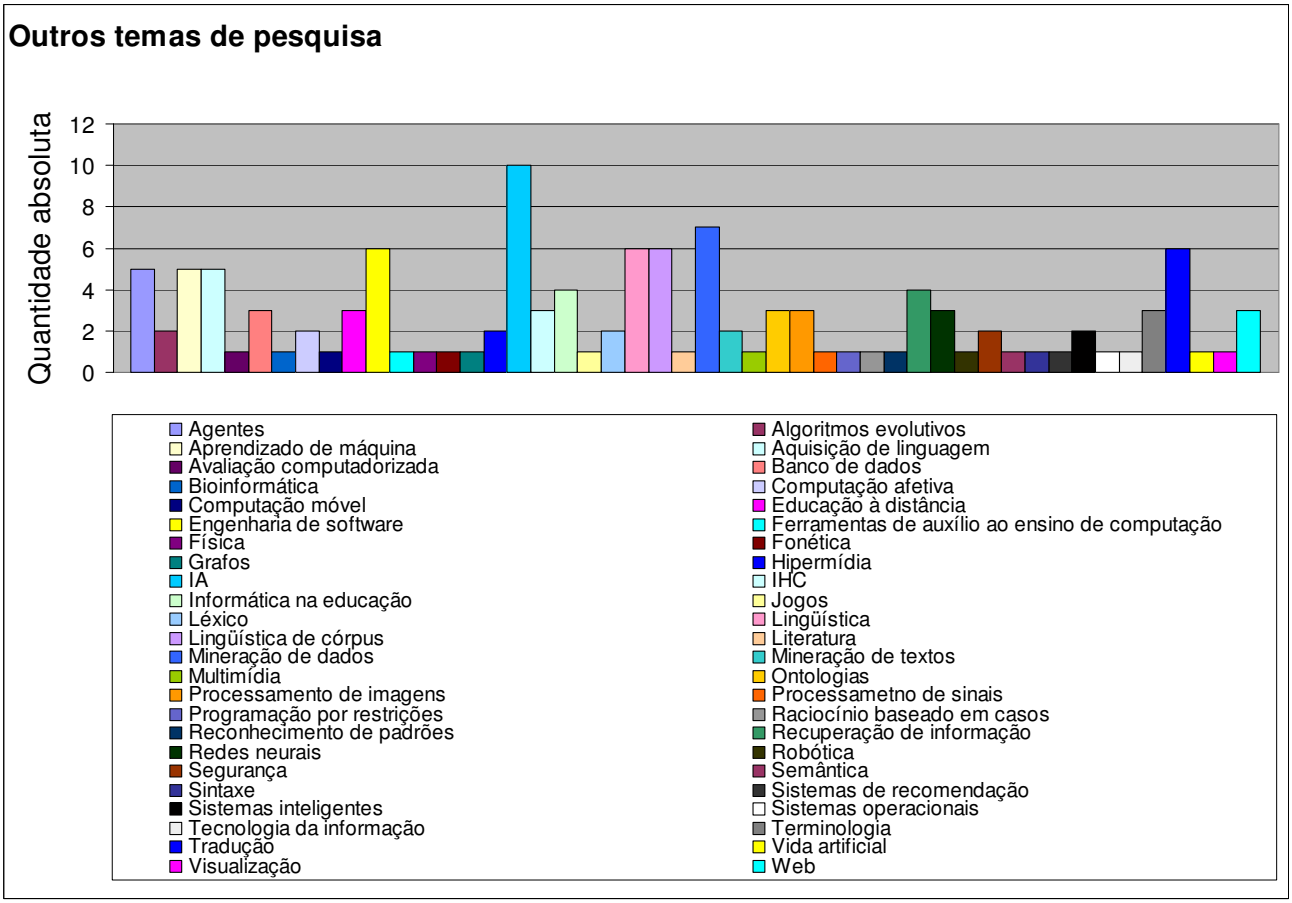

Figura 5. Demais áreas de pesquisa

A Figura 6 mostra a proporção de pesquisadores que são líderes de grupos de pesquisa no CNPq, enquanto a Figura 7 mostra a proporção daqueles que são orientadores em programas de pós-graduação.

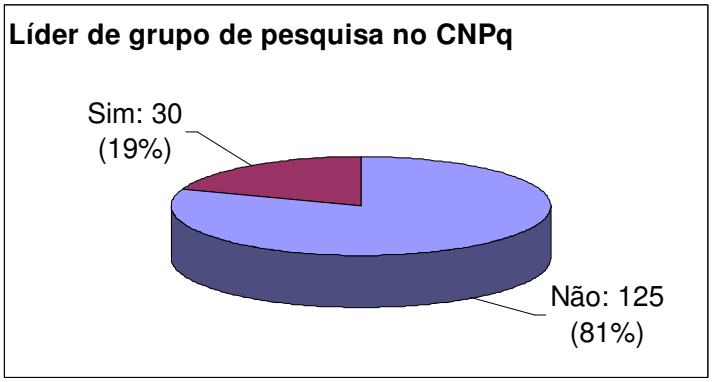

Figura 6. Proporção de pesquisadores que são líderes de grupos de pesquisa no CNPq

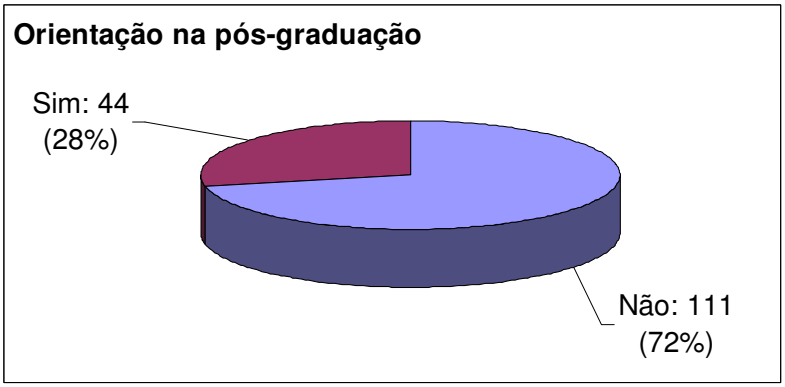

Figura 7. Proporção de pesquisadores que orientam em programas de pós-graduação

A Figura 8 mostra a proporção de pesquisadores que têm projetos financiados, enquanto a Figura 9 exibe o número absoluto de projetos financiados por agências de fomento à pesquisa. Como esperado, pode-se notar que CNPq, FAPESP e CAPES são as agências que mais financiam projetos.

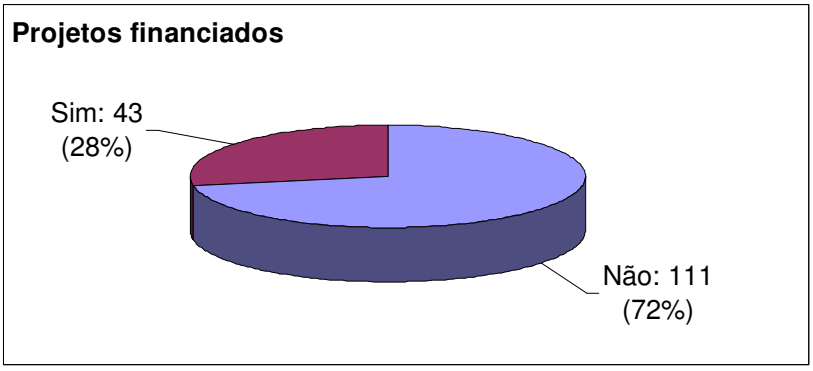

Figura 8. Proporção de pesquisadores com projetos financiados

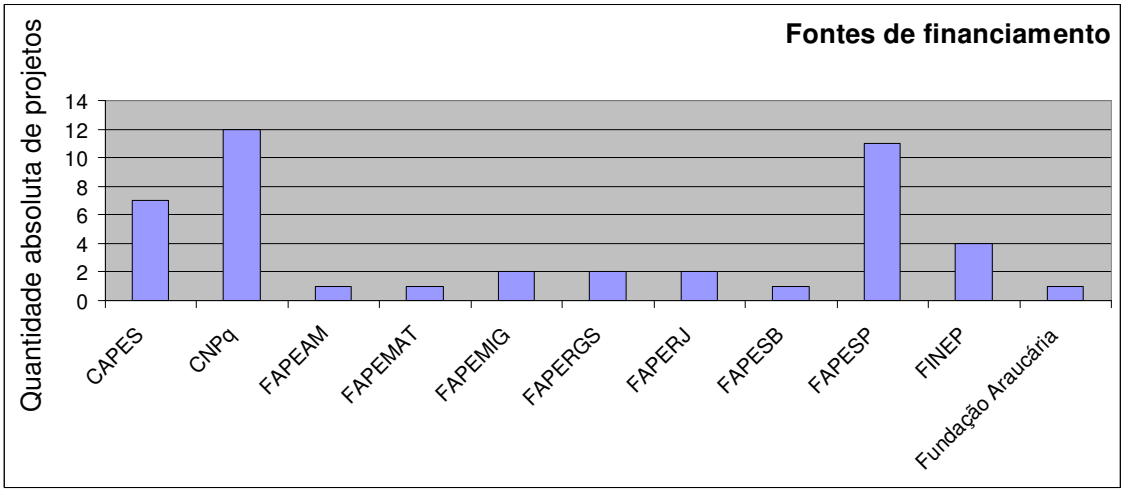

Figura 9. Projetos financiados e agências de fomento à pesquisa

As Figuras 10 e 11 mostram a proporção de pesquisadores que mantêm colaboração com outros grupos de pesquisa nacionais e internacionais, respectivamente. A colaboração internacional ainda é pequena quando comparada com a nacional.

Figura 10. Proporção de pesquisadores com colaboração com outros grupos nacionais

Figura 11. Proporção de pesquisadores com colaboração com grupos internacionais

A Figura 12 mostra a proporção de pesquisadores que são sócios da SBC. Apenas 35% (54 pesquisadores) são sócios da SBC. Naturalmente, isso se atribui ao caráter multidisciplinar da área, em que muitos pesquisadores não são da Computação e, portanto, não fazem parte da SBC.

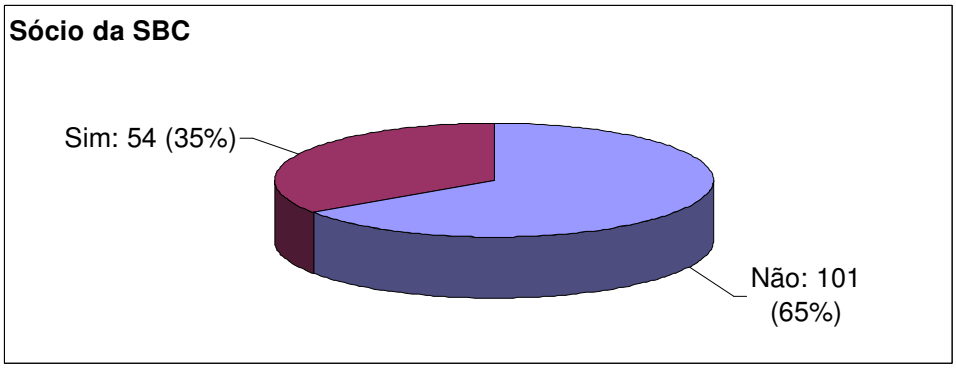

Figura 12. Proporção de sócios da SBC

A Figura 13 exibe a proporção de pesquisadores que são membros da lista de e-mails da CEPLN. Novamente, o número é relativamente pequeno: apenas 62 pesquisadores que responderam à enquete são membros da lista da CEPLN (40%). Isso também quer dizer que, dos atuais 168 membros da lista da CEPLN, apenas 37% responderam à enquete.

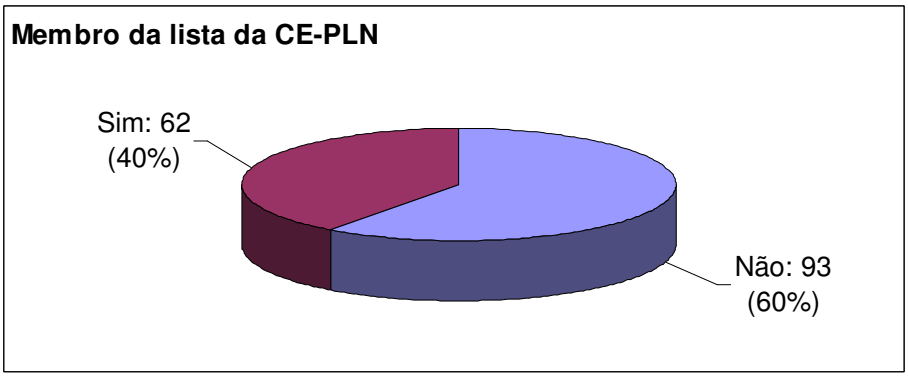

Figura 13. Proporção de pesquisadores membros da CEPLN

A Figura 14 mostra a proporção de pesquisadores que também são sócios de outras sociedades científicas além da SBC. A Figura 15 mostra o número absoluto de pesquisadores em função das outras sociedades a que pertencem. Pode-se ver a predominância da ABRALIN, ACL, ACM, GEL e IEEE: ABRALIN e GEL são sociedades da Lingüística; ACM e IEEE são da Computação; e ACL é multidisciplinar, voltada para ambas as áreas.

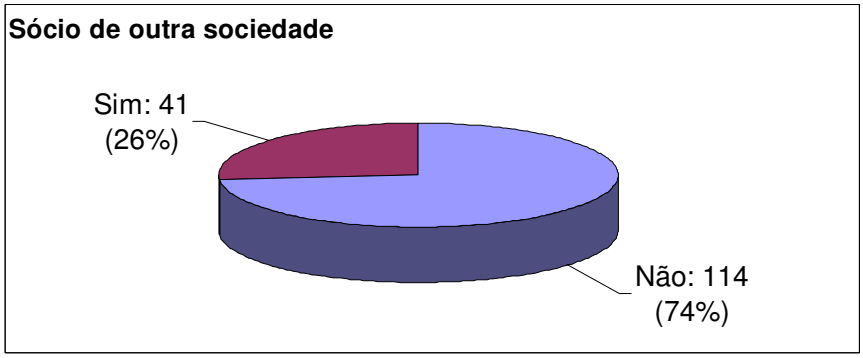

Figura 14. Proporção de pesquisadores sócios de outras sociedades

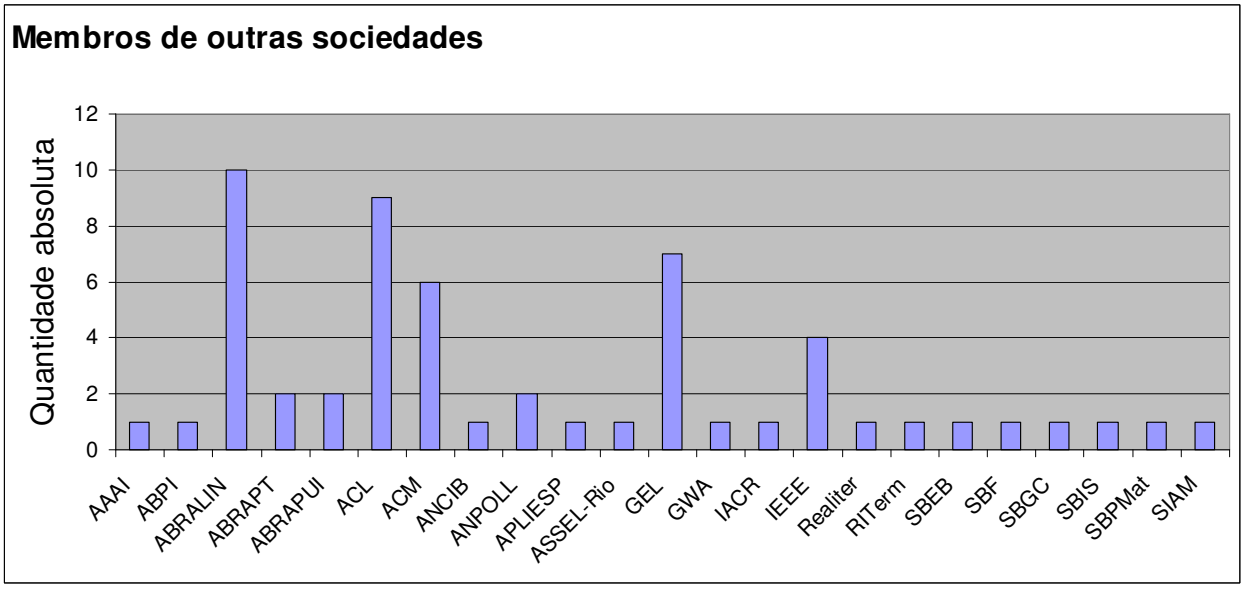

Figura 15. Pesquisadores e outras sociedades

A Figura 16 mostra a distribuição de pesquisadores por estado. Os pesquisadores concentram-se nos estados de São Paulo, Rio Grande do Sul, Paraná e Rio de Janeiro, mas aparecem em 18 estados do país.

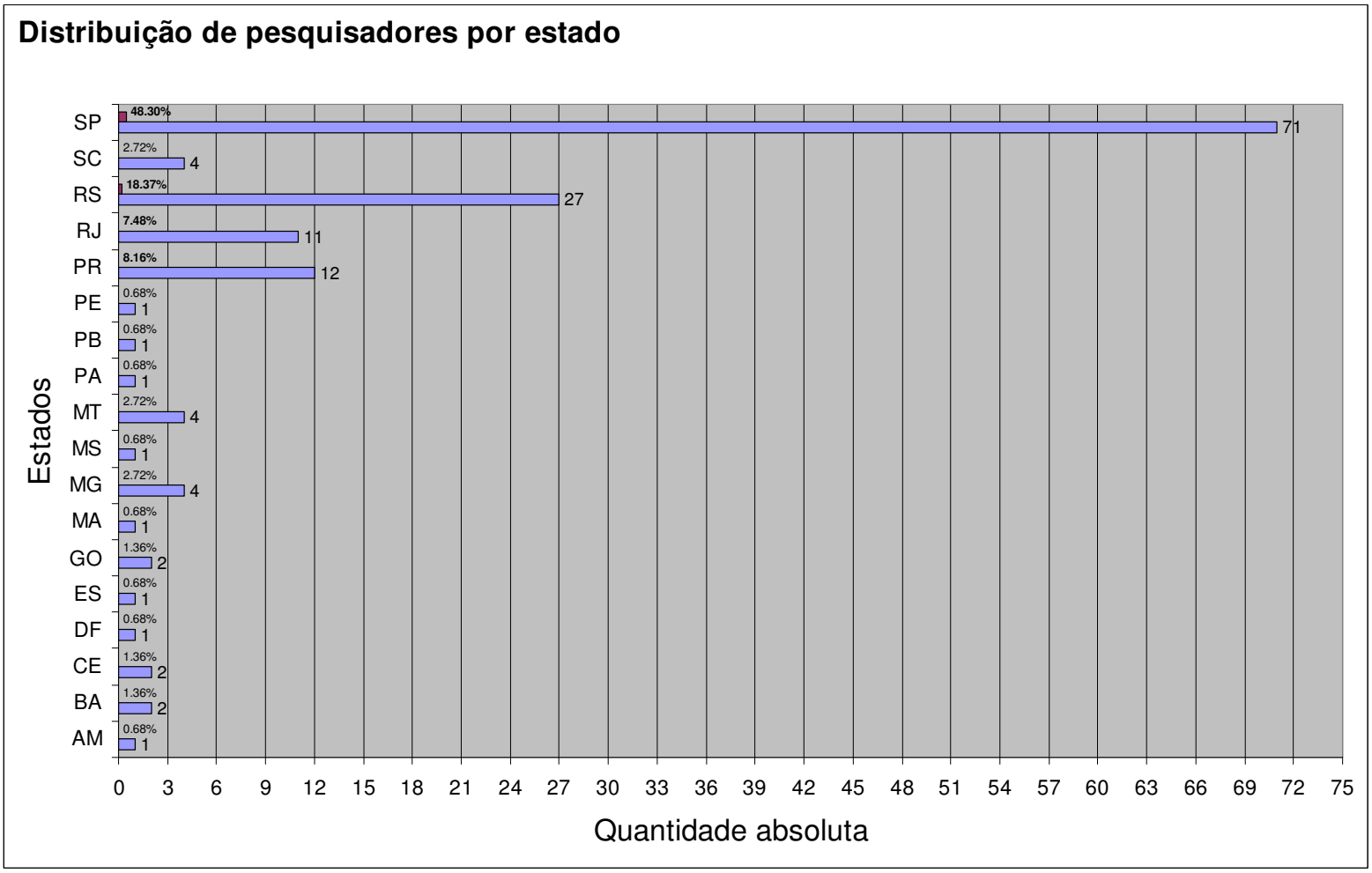

Figura 16. Concentração de pesquisadores por estado

A Figura 17 mostra a distribuição de pesquisadores por área de formação. Pode-se notar a predominância de pesquisadores provenientes da Computação e da Lingüística. A Ciência da Informação, área bastante relacionada, tem participação quase nula no mapeamento realizado.

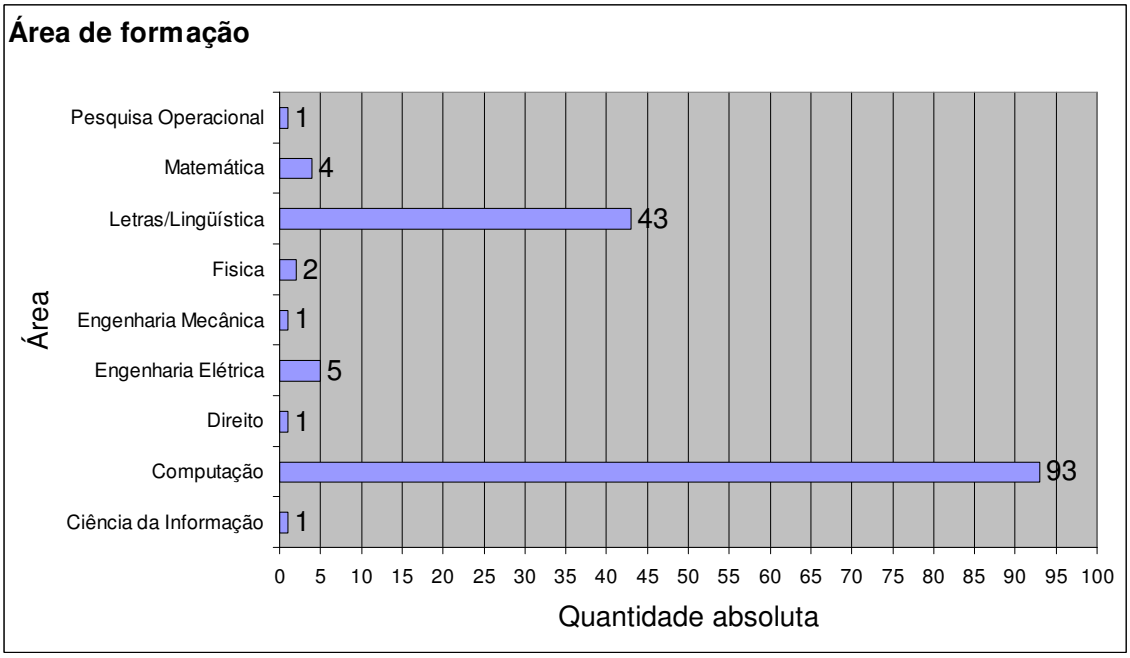

Figura 17. Distribuição de pesquisadores por área de formação

Por fim, a Tabela 1 exibe os desafios atuais da área, mencionados pelos pesquisadores. A tabela indica a quantidade absoluta de vezes e a proporção com que foram citados na enquete. É interessante notar que há desafios de diversas naturezas, já que a enquete propositalmente não especificava que tipo de desafio deveria ser reportado. Alguns desafios são teóricos, outros expressam frustração com relação a alguma área, outros são meramente operacionais, e alguns são secundários em relação à área de PLN.

Tabela 1. Desafios enfrentados pelos pesquisadores

Desafios	**Quantidade**	**Proporção**
Financiamento de projetos	19	14.2%
Ausência de recursos básicos de qualidade para o português (por exemplo, córpus anotados ou não, um bom parser, wordnet, REM)	16	11.9%
Dificuldade em atrair e formar alunos e pesquisadores	9	6.7%
Criação e refinamento de modelos de descrição e análise lingüística	7	5.2%
Montagem e coordenação de esforços multidisciplinares	6	4.5%
Pouca interação entre universidade e empresa nessa área de pesquisa	6	4.5%
Criação de ontologias	5	3.7%
Escassez no país de material de pesquisa relevante (por exemplo, livros de autores renomados da área)	5	3.7%
Interação multidisciplinar	5	3.7%
Anotação de córpus	4	3.0%
Certa marginalização da área tanto na Computação quanto na Lingüística	4	3.0%

Falta de formação computacional básica para lingüistas	4	3.0%
Metodologia de avaliação robusta de recursos, ferramentas e aplicações	3	2.2%
Realizar pesquisa em conjunto com as demais atividades que as universidades demandam	3	2.2%
Divulgação da área e das ferramentas criadas	3	2.2%
Sistematização e automatização das práticas da lexicografia e terminologia	2	1.5%
Resultados insatisfatórios na extração automática de termos	2	1.5%
Maior e melhor interface e interatividade dos sistemas de PLN	2	1.5%
Acesso a bases de dados nacionais e internacionais	2	1.5%
Produção de material de pesquisa em português	2	1.5%
Falta de cooperação entre grupos nacionais	2	1.5%
Pouca integração entre os grupos de pesquisa nacionais e internacionais	1	0.7%
Desenvolvimento de sistemas para aplicações reais e de alto desempenho	1	0.7%
Falta de ações da SBC para favorecer pesquisas multidisciplinares	1	0.7%
Pulverização da pesquisa em subáreas distintas	1	0.7%
Trabalhar com língua portuguesa e ter inserção internacional	1	0.7%
Falta de modelos de processamento integrado dos vários níveis de conhecimento lingüístico	1	0.7%
Desequilíbrio na distribuição de financiamento (grupos já estabelecidos conseguem mais financiamento)	1	0.7%
Criação de um glossário eletrônico	1	0.7%
Lacunas lexicais, culturais e pragmáticas entre inglês e português	1	0.7%
Editor que permita armazenar e manipular os resultados de pesquisas lingüísticas	1	0.7%
Busca de padrões em textos criptografados	1	0.7%
Alinhamento semântico entre línguas naturais	1	0.7%
Resultados insatisfatórios em extração de informação	1	0.7%
Incorporar conhecimento da Lingüística Computacional para construção da web semântica	1	0.7%
Direitos autorais para construção de córpus	1	0.7%
Equipamento computacional ultrapassado	1	0.7%
Poucas pesquisas em Geração de Língua Natural	1	0.7%
Resultados insatisfatórios em recuperação de informação	1	0.7%
Criação de recursos que permitam avanços nas pesquisas em tradução automática	1	0.7%
Poucos avanços recentes na área de tradução automática	1	0.7%
Desenvolvimento de técnicas para anotação automática de dados	1	0.7%
Desenvolvimento de sistemas sem a necessidade de dados anotados	1	0.7%
Pouco desenvolvimento da área de pesquisa	1	0.7%

Os dados obtidos nesse mapeamento também foram cruzados com os dados presentes no Registro de Pesquisadores em Processamento de Línguas Naturais e Lingüística Computacional da América Latina[1], da NAACL (*North American Chapter of the Association for Computational Linguistics*), acessível via web e mantido por Ted Pedersen. Para o cruzamento dos dados, somente os pesquisadores residentes no Brasil foram considerados. 58 pesquisadores que constam na listagem do Registro não responderam à enquete realizada para realização do mapeamento. No total o, o Registro conta com 99 pesquisadores brasileiros cadastrados (residentes no Brasil), sendo que há menos dados disponíveis do que os levantados no presente mapeamento.

A Figura 18 exibe a atividade dos 58 pesquisadores do Registro que não participaram do mapeamento, conforme cadastrado no Registro. É interessante notar que alguns especificaram que são professores, enquanto outros especificaram que são pesquisadores. Acredita-se que essas duas categorias correspondam à categoria professor/pesquisador do mapeamento. Também não há distinção quanto ao nível de formação.

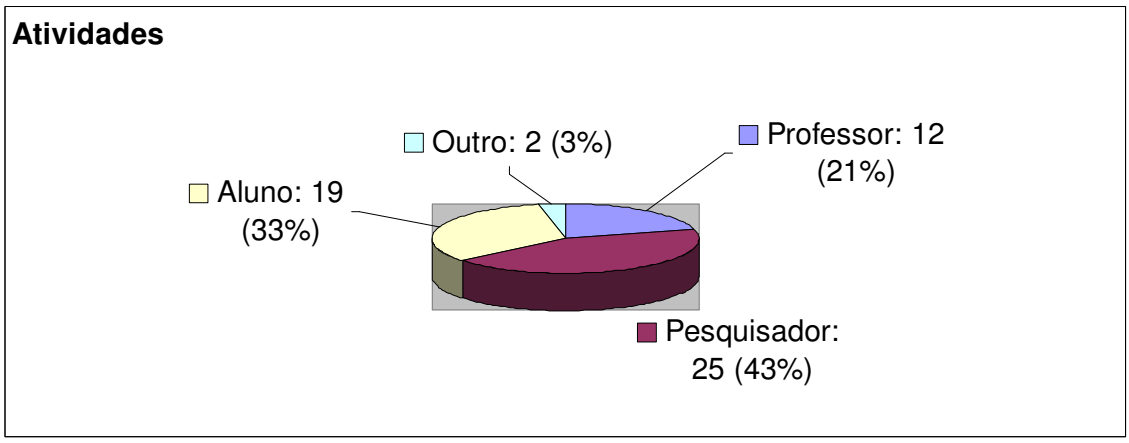

Figura 18. Pesquisadores do Registro e suas atividades

A Figura 19 mostra a distribuição dos 58 pesquisadores do Registro em função de seus temas de pesquisa. Pode-se notar que Lingüística de Corpus e Tradução Automática são os temas que predominam.

[1] http://www.d.umn.edu/~tpederse/registry/registry.cgi

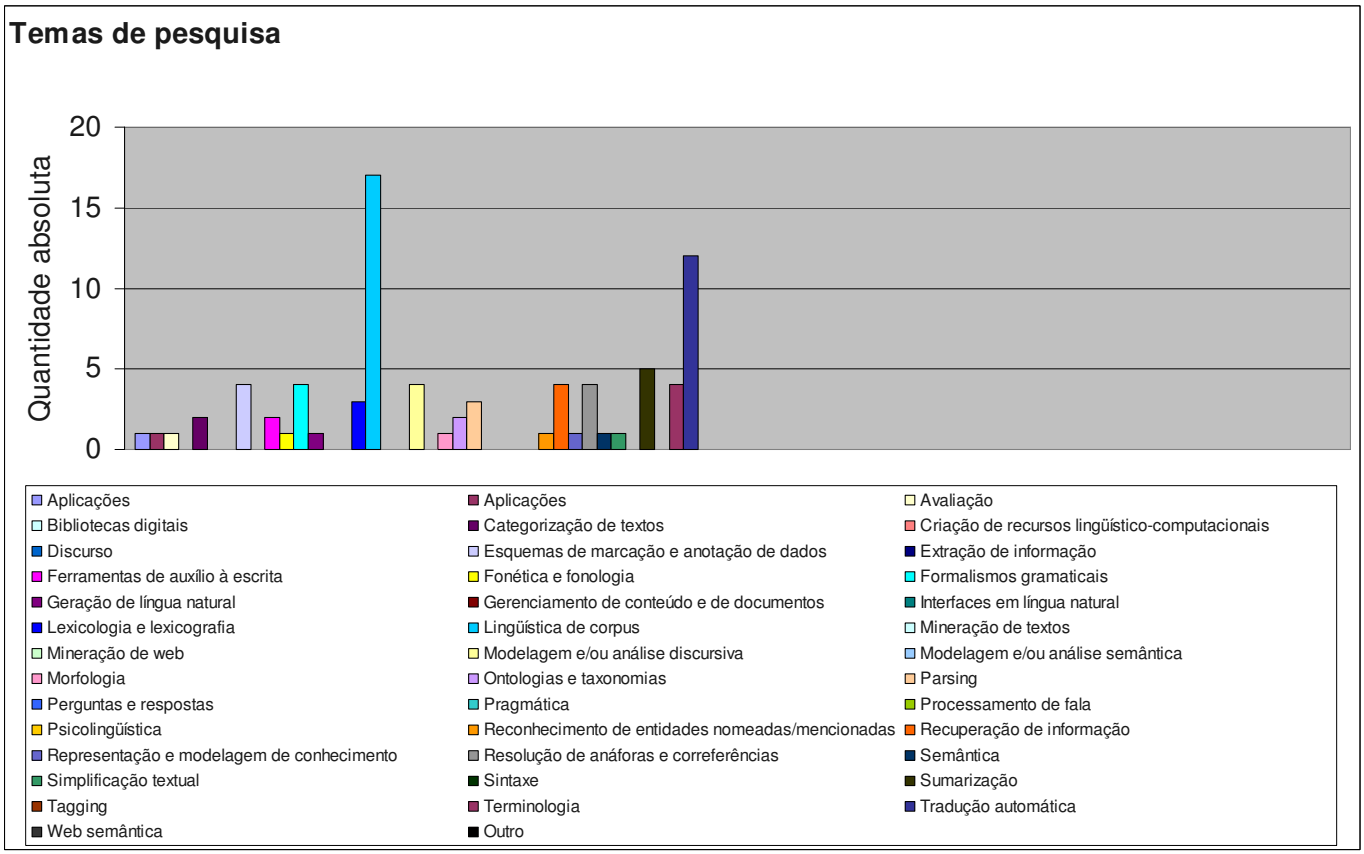

Figura 19. Pesquisadores do Registro e seus temas de pesquisa

A Figura 20 exibe a distribuição dos 58 pesquisadores do Registro por estado. Pode-se perceber que os estados com mais pesquisadores do mapeamento também são os estados com mais pesquisadores do Registro.

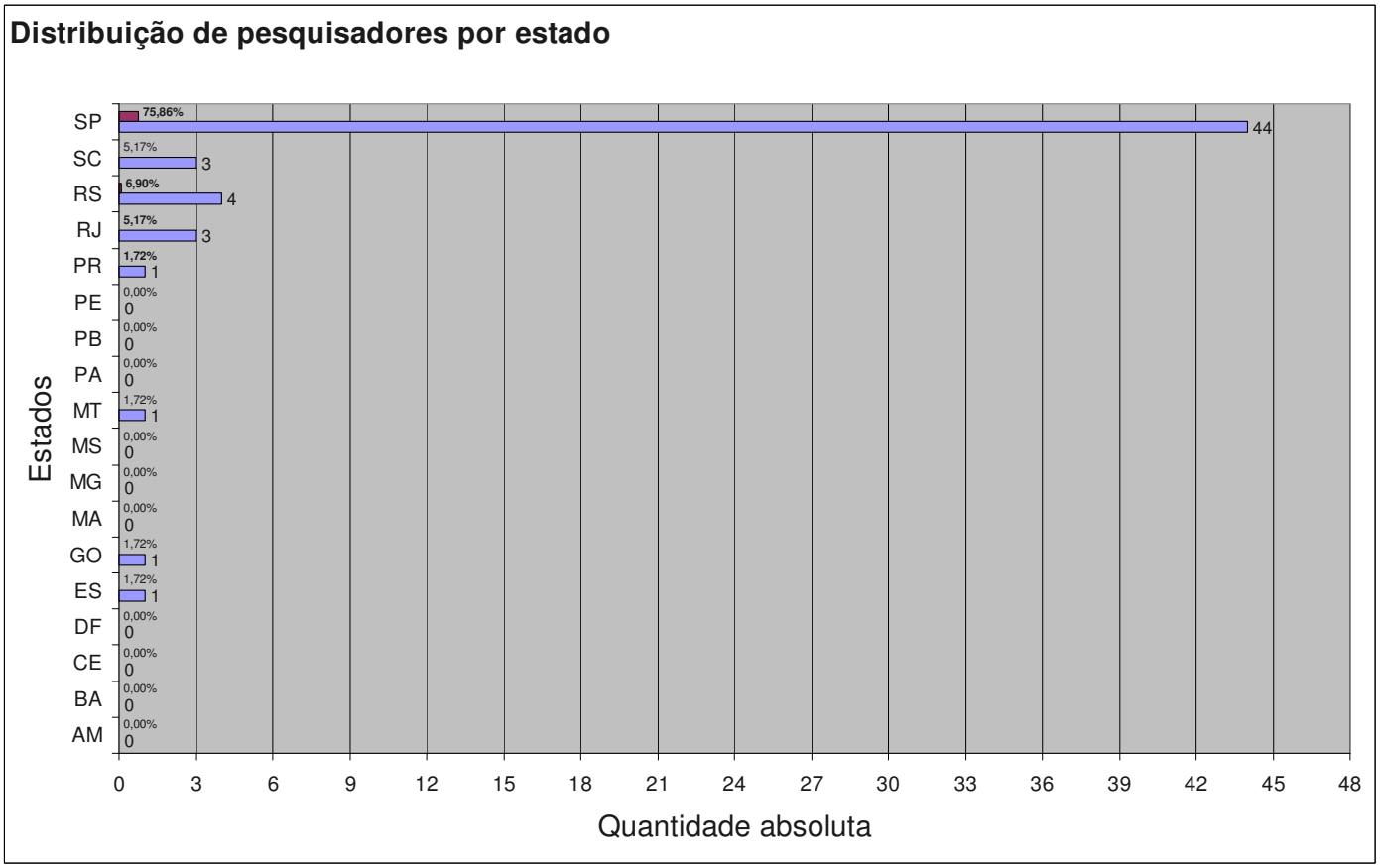

Figura 20. Distribuição dos pesquisadores do Registro por estado

A Figura 21 mostra as outras áreas de pesquisa além de PLN dos 58 pesquisadores do Registro.

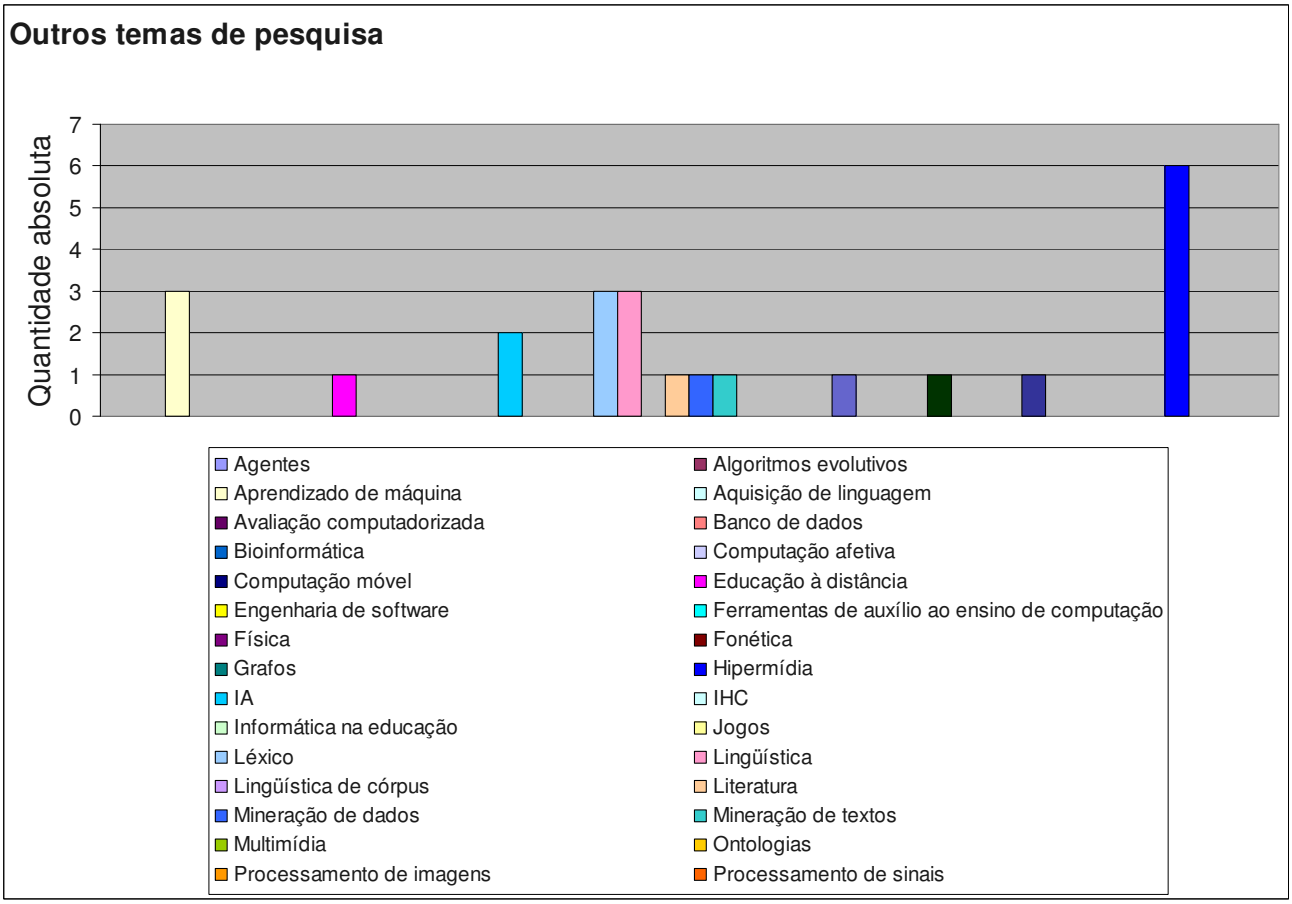

Figura 21. Pesquisadores do Registro e outras áreas de pesquisa

Com base nesses dados, pode-se concluir que o tamanho estimado de pesquisadores que trabalham direta ou indiretamente em PLN é de no mínimo 254, aproximadamente, como demonstra a Tabela 2. Certamente há pesquisadores que estão no Registro e que não participaram da enquete e não são membros da lista da CEPLN, o que aumentaria um pouco mais o número de interessados em PLN. Entretanto, no momento em que este mapeamento foi feito, não se tinha em mãos essa informação. Por essa razão, os 58 pesquisadores do Registro não foram levados em conta nessa estimativa.

Tabela 2. Pesquisadores com interesse em PLN

Fonte	**Quantidade absoluta**
Pesquisadores que responderam a enquete	148
Membros da lista da CEPLN que não participaram da enquete	106
Total	*254*

3. Considerações Finais

Pelo que se sabe, esse é o primeiro mapeamento detalhado da área de PLN no Brasil. Espera-se que, com os dados obtidos, haja informação suficiente para direcionar futuras ações na área.

Os dados apresentados foram pouco refinados ou filtrados. Eles são apresentados praticamente na forma como aparecem na enquete realizada. Por essa razão, é interessante notar que algumas áreas muitas vezes são consideradas como fazendo parte de PLN e, outras vezes, não. É o caso, por exemplo, da Lingüística de Corpus. Outras curiosidades são: áreas como Aprendizado de Máquina aparecem desvinculadas de Inteligência Artificial (devido ao fato da enquete não fazer tipo algum de restrição ou exigência quanto à especificidade dos temas), a participação pequena, mas mais expressiva do que outras áreas, da Matemática e da Engenharia Elétrica (esta, em particular, tem vários trabalhos em Processamento de Fala, área de grande correlação com PLN), dentre outras.

Em função da experiência e vivência dos autores deste relato na área de PLN, acredita-se que os resultados apresentados aqui refletem bem a área. Entretanto, como se pode perceber, a enquete que deu origem a esse mapeamento não foi respondida por todos que estão ligados de alguma forma à área de PLN. Há muitos membros da CEPLN e do Registro de Pesquisadores em Processamento de Línguas Naturais e Lingüística Computacional da América Latina que não participaram. Iniciativas futuras devem tentar obter maior abrangência.

Apêndice A – Reprodução da enquete realizada

Mapeamento da comunidade de Processamento de Línguas Naturais e Lingüística Computacional no Brasil

A Comissão Especial de Processamento de Linguagem Natural da Sociedade Brasileira de Computação (CEPLN-SBC - www.sbc.org.br/cepln) convida a comunidade brasileira (de qualquer área de conhecimento) que trabalha direta ou indiretamente com Processamento de Línguas Naturais e/ou Lingüística Computacional para participar da pesquisa abaixo para mapeamento da área no Brasil.

Esse mapeamento tem o objetivo de realizar um levantamento da área e seus pesquisadores e de subsidiar ações no Brasil.

*Obrigatório

Nome completo (sem abreviações)
*

Cidade em que reside
*

Estado
*

E-mail
*

Nome completo (sem abreviações) da universidade / instituição / empresa a qual está vinculado

Atividade / função *

- Professor e/ou pesquisador com doutorado finalizado
- Professor e/ou pesquisador com mestrado finalizado (sem doutorado finalizado)
- Professor e/ou pesquisador sem doutorado e sem mestrado
- Aluno de doutorado
- Aluno de mestrado
- Aluno de graduação
- Outro:

Área de formação *Especifique: Computação, Letras/Lingüística, Ciência da Informação, Engenharia Elétrica, Matemática, Estatística, Filosofia, etc.

Se faz ou já fez graduação, especifique a universidade / instituição da graduação. Escreva o nome completo da universidade / instituição (sem abreviações)

Se faz ou já fez mestrado, especifique a universidade / instituição do mestrado. Escreva o nome completo da universidade / instituição (sem abreviações)

Se faz ou já fez doutorado, especifique a universidade / instituição do doutorado. Escreva o nome completo da universidade / instituição (sem abreviações)

Temas que investiga (direta ou indiretamente) na área de Processamento de Línguas Naturais / Lingüística Computacional*
Marque todas as opções que se aplicam

- [ ] Interpretação de língua natural
- [ ] Geração de língua natural
- [ ] Processamento de fala
- [ ] Fonética e fonologia
- [ ] Morfologia
- [ ] Sintaxe
- [ ] Semântica
- [ ] Discurso
- [ ] Pragmática
- [ ] Lexicologia e lexicografia
- [ ] Terminologia
- [ ] Ontologias e taxonomias
- [ ] Formalismos gramaticais
- [ ] Esquemas de marcação e anotação de dados
- [ ] Criação de recursos lingüístico-computacionais
- [ ] Representação e modelagem de conhecimento
- [ ] Tagging
- [ ] Parsing
- [ ] Reconhecimento de entidades nomeadas/mencionadas
- [ ] Modelagem e/ou análise semântica
- [ ] Modelagem e/ou análise discursiva
- [ ] Resolução de anáforas e correferências
- [ ] Ferramentas de auxílio à escrita
- [ ] Sumarização
- [ ] Tradução automática

- [ ] Simplificação textual
- [ ] Categorização de textos
- [ ] Perguntas e respostas
- [ ] Extração de informação
- [ ] Recuperação de informação
- [ ] Mineração de textos
- [ ] Mineração de web
- [ ] Gerenciamento de conteúdo e de documentos
- [ ] Web semântica
- [ ] Bibliotecas digitais
- [ ] Lingüística de corpus
- [ ] Psicolingüística
- [ ] Interfaces em língua natural
- [ ] Aplicações
- [ ] Avaliação
- [ ] Outro:

Sua área de pesquisa principal é Processamento de Línguas Naturais / Lingüística Computacional? *

- ○ Não
- ○ Sim

Também faz pesquisa em outras áreas? *

- ○ Não
- ○ Sim

Se a resposta anterior foi "sim", especifique em quais outras áreas desenvolve pesquisa

É líder de algum grupo de pesquisa no CNPq? *

- ○ Não
- ○ Sim

Se a resposta anterior foi "sim", especifique o ano de formação do(s) grupo(s)

É orientador em algum programa de pós-graduação? *

- Não
- Sim

Tem projetos financiados? *

- Não
- Sim

Se a resposta anterior foi "sim", especifique as agências / fontes de financiamento. Se possível, especifique também o número de projetos financiados por cada agência / fonte e os temas dos projetos, assim como número de alunos envolvidos, páginas web dos projetos (se houver) e outras informações que julgar interessantes.

Tem cooperação com outros grupos de pesquisa nacionais? *

- Não
- Sim

Tem cooperação com grupos de pesquisa internacionais? *

- Não
- Sim

É sócio da Sociedade Brasileira de Computação (SBC)? *

- Não
- Sim

Está associado à lista de discussão da Comissão Especial de Processamento de Linguagem Natural da SBC? *

- Não
- Sim

É sócio de alguma outra sociedade (nacional ou internacional) além da SBC? *

- Não
- Sim

Se a resposta anterior foi "sim", especifique a quais outras sociedades está associado

Quais são os desafios que encontra ou que prevê em seu trabalho na área? Os desafios podem ser de qualquer natureza: práticos e/ou operacionais, de pesquisa (teóricos e/ou práticos), de financiamento, etc.

Enviar