LADEIRA, Ana Paula. Processamento de linguagem natural: caracterização da produção científica dos pesquisadores brasileiros. 259f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

Sinais evidentes de contribuições de grandes campos disciplinares marcaram e têm influenciado fortemente as pesquisas na área de processamento de linguagem natural (PLN), dentre eles a ciência da computação, a ciência da informação e a linguística. Sendo assim, a presente tese pretendeu utilizar o conhecimento acumulado ao longo dos últimos 40 anos em PLN e revelado no ARIST, como referência para selecionar e analisar a produção científica da comunidade acadêmica nacional da área. As publicações nacionais foram coletadas automaticamente da Plataforma Lattes, e um instrumento de seleção automática foi construído a partir da análise de assunto dos artigos de revisão do ARIST. Este instrumento foi utilizado para selecionar, de maneira automática, as publicações nacionais atinentes para a área de PLN. Dentre as 621 publicações consideradas da área, definiu-se o material empírico, constituído por uma amostra de 68 trabalhos, que foi submetido à análise de conteúdo. Essa análise permitiu elucidar as temáticas discutidas pela comunidade científica nacional. Ao analisar todas as publicações atinentes para a área de PLN, observou-se que a grande maioria da produção científica foi publicada depois do ano 2.000. Além disso, a participação da ciência da informação tem sido muito modesta, sendo que a ciência da computação e a linguística foram responsáveis por quase 85% da produção nacional. Doze pesquisadores foram responsáveis por mais de 20% de toda a produção nacional, sendo que dentre eles, nove são da ciência da computação, dois da linguística, e um é da engenharia elétrica. Além disso, vale destacar que dentre esses doze pesquisadores, sete fazem parte do grupo de pesquisa NILC. Dentre as problemáticas mais discutidas, foi possível observar que: a tradução foi intensamente abordada na década de 90; os estudos com indexação diminuíram a partir da década de 80; e que as pesquisas sobre classificação passaram por um período de dormência na década de 90; e que existe uma tendência clara na área de PLN de desenvolvimento de pesquisas em sumarização automática. Outro aspecto que a pesquisa revelou foi que a ciência da informação tem priorizado as pesquisas em indexação automática,

seguido da análise de conteúdo, enquanto que a ciência da computação tem priorizado as pesquisas em tradução e sumarização. A análise de conteúdo realizada nas 68 publicações selecionadas permitiu revelou que a recuperação de informação foi a problemática que teve maior destaque na produção científica nacional. Dos trabalhos analisados sobre sumarização, observou-se que somente dois usaram a abordagem profunda e produziram sumários, e que a maioria das pesquisas em sumarização automática tem privilegiado a abordagem empírica (para gerar extratos). As pesquisas em tradução automática têm utilizados métodos estatísticos e regras de transferências, com resultados muito próximos. Apesar das pesquisas em PLN estarem ocorrendo em campos disciplinares diferentes da ciência da informação, os estudos realizados precisam ser conhecidos, pois esta última pode se beneficiar das ferramentas computacionais desenvolvidas, aplicando-as em processos clássicos inerentes ao campo, tais como catalogação, recuperação e representação de informação.