Compilação de *corpus* especializado sobre o contexto de doenças e pragas do cafeeiro

Henrique de Almeida Regitano[1]
Leandro Henrique Mendonça de Oliveira[2]

As doenças e pragas que incidem sobre os cafeeiros causam redução da produtividade e da qualidade do café colhido, além de elevarem os custos de produção e os riscos ambientais advindos da aplicação de medidas de controle. O uso indiscriminado de agroquímicos também induz a resistência das pragas e doenças aos agentes de controle. Para melhor apropriação do conhecimento que subjaz a todo esse cenário, é imprescindível a sistematização da terminologia envolvida, afinal, é por intermédio dos termos que se veicula conhecimento especializado. Assim, por meio de uma parceria com o Grupo de Estudos e Pesquisa em Terminologia (GETerm), situado no Departamento de Letras da Universidade Federal de São Carlos (UFSCar), Campus São Carlos, em modalidade de estágio, este trabalho apresenta a parte inicial da organização da informação e representação do conhecimento no contexto de doenças e pragas do cafeeiro, tendo como objetivo a compilação e a limpeza de um *corpus* textual especializado sobre este domínio do conhecimento, para que seja trabalhado terminologicamente.

A seleção dos textos pertinentes à composição do *corpus* foi feita seguindo critérios definidos no início do trabalho. Esses critérios delimitam características dos textos, como o tipo textual, sendo este tipo o de resumo expandido; sua fonte, o Simpósio de Pesquisa dos Cafés dos Brasil (SPCB) em suas edições entre os anos 2000 e 2013; e o assunto abordado, as doenças e as pragas do café. Devido ao fato de os textos serem disponibilizados em formato PDF, é necessário que sejam convertidos em arquivos de texto puro para que possam ser analisados por computador. Para essa tarefa, foi utilizado o programa *ABBYY PDF Transformer 3.0*, devido à qualidade

[1] Universidade Federal de São Carlos (UFSCar)
[2] Embrapa Informática Agropecuária

superior do arquivo convertido. A limpeza do *corpus* foi realizada utilizando o editor de textos *Kate*, em ambiente Linux. Essa escolha foi baseada na necessidade de remoção de tabelas e figuras dos textos, e devido à sua capacidade de detecção de expressões regulares. A pesquisa pelos textos foi realizada ano a ano, percorrendo manualmente o website do Sistema Brasileiro de Informação do Café (SBICafé), onde se encontram os resumos expandidos de trabalhos apresentados no SPCB dos anos referidos, e nos anais do SPCB disponíveis no website do Consórcio Pesquisa Café, buscando indícios de abordagem do tema nos títulos dos trabalhos. Foi feito o download dos arquivos PDF individuais de cada texto, sendo armazenados separadamente de acordo com o ano de apresentação no SPCB. Cada arquivo foi convertido individualmente em arquivos de texto puro, preservando os originais para uso comparativo durante a limpeza, evitando a perda de informações importantes dos textos durante esse processo.

A partir dessa pesquisa, foi compilado um *corpus* textual especializado de tamanho pequeno, contendo menos de quinhentos textos e em torno de um milhão de palavras contadas antes do início da limpeza.

Palavras-chave: Pragas e doenças do cafeeiro, *corpus* textual, conhecimento especializado.