1 O dadosSemiotica

O dadosSemiotica (dS) é um software livre que provê uma interface online para a organização e realização de análises de textos verbais em língua portuguesa.

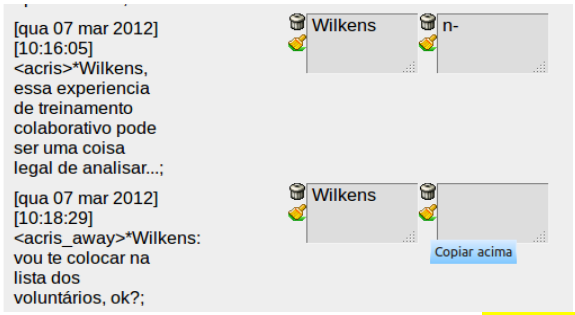

Ilustração 1: interface de análises manuais PRECISA ATUALIZAR QUANDO SAIR A VERSÃO 2.0

A interface de análises manuais permite visualizar a sequência de sentenças (mostra 50 sentenças por página PRECISA ATUALIZAR QUANDO SAIR A VERSÃO 2.0) e repetir, com um clique, análises que abrangem mais de uma sentença. O programa possui várias ferramentas para agilizar o processo de análise do texto, descritas no manual do dS. Após a realização das análises, permite recuperar tabelas com os resultados das análises manuais e automáticas e obter algumas estatísticas descritivas do *corpus*. É indicado principalmente para quem deseja trabalhar com corpora de grandes dimensões, mas pode ser usado em pesquisas mais específicas, com ótimo aproveitamento para pesquisas interdisciplinares, inclusive no limiar entre as Ciências Humanas e as Ciências Exatas.

O software foi desenvolvido por uma equipe de desenvolvedores do grupo Texto Livre: Semiótica e Tecnologia, registrado no CNPq[1]. A equipe inicial, responsável vela versão 1.0, era formada por Ana Matte, coordenadora acadêmica do grupo Texto Livre, Rubens Ribeiro, desenvolvedor do framework SIMP no qual o dadosSemiotica foi desenvolvido, William Colen, desenvolvedor do corretor gramatical do OpenOffice para o português brasileiro, e Hugo Leonardo Canalli, coordenador de desenvolvimento de aplicativos do Texto Livre. O lançamento da versão 1.0 aconteceu em 27/07/2012, durante o WSL2012 – Workshop Internacional de Software Livre – durante o fisl13, em Porto Alegre (Matte, Takiguti, Colen, Canalli, 2012)[2].

O desenvolvimento da versão 2.0 aconteceu em duas etapas: a primeira, contando com a mesma equipe que desenvolveu a versão 1.0, resultou em inovações, como novas funcionalidades e aperfeiçoamento do design, a que chamamos de versão 1.5, pois não chegou a ser concluída antes de acontecerem grandes alterações na equipe. Na segunda etapa, iniciou-se uma parceria com a empresa de desenvolvimento de software Conexum, mais especificamente na pessoa do desenvolvedor especialista na área de Inteligência Artificial e processamento de Línguas Naturais,

1 Página do Grupo http://textolivre.org e, no CNPq, http://dgp.cnpq.br/dgp/espelhogrupo/4246802692010460.
2 Apoio: CNPq (Processo N° 310304/2012-1) e FAPEMIG (Processo N° PPM-00206-10).

Daniel Nehme Müller, para concluir o desenvolvimento do dadosSemiotica[3], desta vez já com funcionalidades orientadas pela pesquisa em Categorias Fechadas[4].

O dadosSemiotica, embora tenha sido concebido para realização de análises de semiótica de linha francesa, pode ser utilizado para pesquisas em outras teorias, especialmente aquelas que permitam trabalhar com a sentença como unidade mínima, pois o usuário pode definir o conjunto de categorias com as quais deseja trabalhar. Foi concebido para agilizar pesquisas interdisciplinares e também possui, com o Módulo de Semiótica, suporte para uma utilização didática no ensino da teoria.

O programa, durante o *upload* do texto-objeto, possui um módulo de preprocessamento morfossintático que divide o texto de entrada em sentenças e guarda os resultados completos de análises morfossintáticas automáticas, as quais usam o motor do Corretor Gramatical do OpenOffice (CoGrOO)*. Atualmente, também conta com um módulo de pré-processamento de chat compatível com o software de chat Konversation[5] (*CITAR OS NOVOS APLICATIVOS COMPATÍVEIS), que recolhe informações de chat, tais como nick do falante, mudanças de apelido, entradas e saídas, e prepara o chat para ser processado pelo CoGrOO. *PRECISA ATUALIZAR QUANDO SAIR A VERSÃO 2.0

Para pesquisas científicas, recomenda-se dedicar uma instalação do programa para uso de um grupo específico, pois como as categorias são criadas pelos usuários, a mistura de diferentes categorias de análise, provenientes de diferentes teorias e objetivos, pode diminuir sensivelmente a usabilidade do programa, em virtude de homônimos nos nomes das categorias e número excessivo de categoriasPRECISA ATUALIZAR QUANDO SAIR A VERSÃO 2.0. A instalação pode rodar em localhost (com um servidor rodando em seu computador), mas como as análises morfossintáticas são feitas via web, uma conexão com a internet é requerida para upload mesmo nesse caso, em virtude do processamento morfossintático realizado a distância.

Testes da versão em desenvolvimento podem ser realizados na interface aberta: <LINK A DEFINIR*> e a versão 2.x*, da qual trata este artigo, pode ser baixada na página do projeto: <LINK A DEFINIR*>. Outras informações – até o presente momento restritas à versão 1.0* – podem ser obtidas na página do Grupo de Pesquisa (http://textolivre.org/site/softwares-do-texto-livre/dadosSemiotica/), na qual outras informações estão disponíveis, incluindo o manual do usuário.*PRECISA ATUALIZAR QUANDO SAIR A VERSÃO 2.0

3 Apoio: FAPEMIG (Processo N° CHE – PPM-00260-16).

4 Apoio: CNPq (Processo N° 305937/2015-4).

5 Konversation é um software cliente de IRC (Internet Relay Chat), um antigo. O IRC, datado de 1988 e bastante usado ainda, especialmente por comunidades de Software Livre, é um protocolo de conversas instantâneas pela internet (https://konversation.kde.org/). O Konversation possui interface amigável e permite registrar as conversas automaticamente em logs, cuja estrutura é a base do parser criado para importação de chat no dadosSemiotica v.1.0, alocando informações como tempo entre cada conversa, nick do destinador, mudanças de nick, dentre outras.