4 O que é um projeto dS?

No dadosSemiotica, ou dS, **executar um projeto é fazer análises de um conjunto de textos para um conjunto de categorias.** Por mais simples que pareça, é importante observar algumas estratégias para tirar o proveito máximo do programa.

- Um projeto no dadosSemiotica (doravante projeto dS) é um conjunto de textos a ser analisado sob um conjunto de categorias teóricas de análise.

É possível acrescentar textos e categorias a qualquer momento. Embora isso seja uma vantagem técnica, para o analista é necessário muito cuidado com esse tipo de decisão, especialmente ao optar por acrescentar textos num projeto cuja análise já foi iniciada, pois dessa alternativa pode decorrer uma não homogeneidade na quantidade de análises em cada texto do *corpus*. Caso uma análise não seja realizada para um texto ou uma categoria, os resultados são marcados como nulos na tabela final, mas o acréscimo de categorias é menos problemático, pois escolher uma categoria implica necessariamente optar por realizar sua análise para todos os textos do projeto.

- Assim, **sugerimos que o projeto dS seja definido a partir do conjunto de textos**; se for usar um conjunto diferente, mesmo que haja intersecção entre os conjuntos de textos de algum projeto existente, a criação de um novo projeto é mais indicada pois favorece resultados mais robustos[1].

É importante lembrar que a inclusão de textos e categorias no sistema correspondem, respectivamente, às etapas de coletas de dados e estudo teórico-metodológico, as quais precedem a análise. Por esse motivo, o usuário que inclui esses dados no sistema não é o analista, é o gerente.

Se você não é o gerente da instalação do dadosSemiotica que está usando, deverá encaminhar ao gerente o pedido de inclusão desses dados antes de criar seu projeto; mesmo nesse caso, siga as instruções sobre registro de textos e categorias para fazer o pedido de forma adequada.

- Em primeiro lugar, devemos ter em mente que o projeto dS pode ou não ser do tamanho do seu projeto de pesquisa.

Dependendo das proporções e objetivos da pesquisa, pode ser que uma pesquisa corresponda a muitos projetos dS; por outro lado, um mesmo projeto dS pode ser utilizado em diferentes pesquisas, principalmente se sua criação envolveu uma reflexão prévia acerca de seu escopo teórico e metodologia, provendo-lhe a consistência necessária. É por isso que, para criar o projeto no dadosSemiotica (o projeto dS), em primeiro lugar você deve pensar se o corpus da sua pesquisa será analisado todo da mesma forma ou será subdividido para análises diferentes. A partir disso, criará tantos projetos dS quantos for necessário. **Dê um nome ao projeto dS que ajude a lembrar** (mesmo muitos anos depois) qual era o escopo daquele projeto específico.

1 NA VERSÃO 1.5 ainda não é possível ao analista copiar análises de um projeto para outro. NA VERSÃO 1.5 também ainda não é possível realizar recuperação de dados como usuário gerente, de modo que não é possível cruzar dados de analistas diferentes. Estas funcionalidades estão na lista de melhorias previstas para o futuro.

- Escolha subdividir ou não seu trabalho em um ou mais projetos dS conforme as necessidades de seu projeto principal.

Se, em sua pesquisa, você estiver testando diferentes formas de abordar um determinado aspecto teórico, você poderia optar por criar um projeto dS diferente para cada abordagem ou criar categorias diferentes para o mesmo aspecto teórico especificando a abordagem no nome, num único projeto dS. A diferença fundamental é que, no primeiro caso, não poderá comparar diretamente os resultados (colocá-los na mesma tabela de saída), pois somente é possível obter resultados dentro de um mesmo projeto dS[2]. Mas, como é possível realizar as análises de cada categoria com total independência na interface de análises do dadosSemiotica e como é possível obter resultados parciais, é possível realizar esse trabalho comparativo num único projeto sem qualquer prejuízo metodológico.

- Divida o texto conforme a necessidade do seu projeto; o dadosSemiotica proverá outras subdivisões automaticamente, tendo em vista a organização das análises.

O dadosSemiotica divide o texto de duas formas: primeiro, cada linha de um *arquivo txt*[3] corresponde a uma entrada, a que chamaremos de parágrafos. Cada parágrafo é enviado ao módulo de pré-processamento morfossintático, o qual subdivide o parágrafo em sentenças, que são a unidade mínima de análise. O tamanho de cada sentença depende de sua estrutura linguística e por isso essa análise só é compatível com textos em língua portuguesa, pois o módulo de pré-processamento morfossintático é baseado no CoGrOO (Corretor Gramatical do OpenOffice para a língua portuguesa do Brasil). Na VERSÃO 1.0 do dadosSemiotica não é possível ainda excluir essa análise do pré-processamento. Se você deseja uma divisão em trechos menores, separe cada trecho por linha no arquivo de entrada antes do upload, como no exemplo da figura 2 em que foi dado como entrada um texto dividido por segmentos delimitados de vogal a vogal.

2 Essa limitação acontece nas versões 1.0 e 1.5, mas está em desenvolvimento a abertura dessa possibilidade de cruzamento entre projetos de um mesmo analista – pelo próprio – ou entre projetos de diferentes analistas, pelo usuário gerente.

3 O arquivo em txt é um arquivo de texto plano, sem formatação, que permite poucas marcas não textuais. Dentre elas, a quebra de linha e o adentramento. Um parágrafo, por exemplo, mesmo que ocupe várias linhas na tela do computador, corresponde a uma única linha do arquivo.

Ilustração 2: Texto de entrada usando escrita fonológica e dividido pela introdução de uma quebra de linha após cada segmento. O CoGrOO não reconhece as sentenças e mantém a divisão por linhas para as sentenças.

A divisão feita pelo CoGrOO é precedida pela divisão em parágrafos (linhas, figura 3).

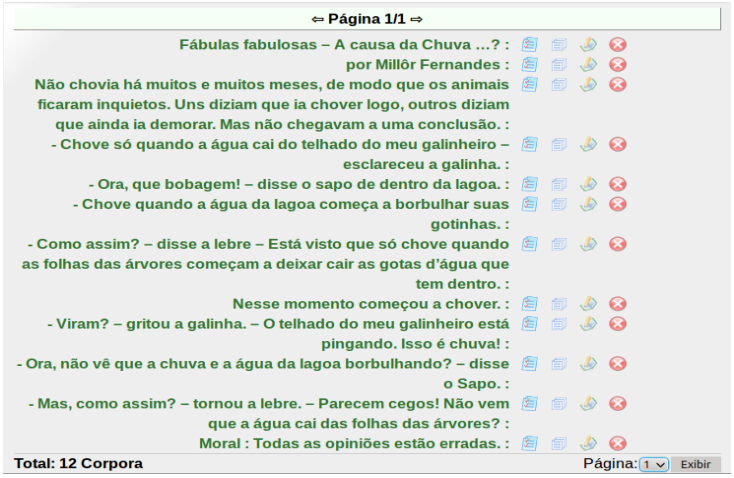

Ilustração 3: Parágrafos do texto de entrada "A causa da Chuva", de Millor Fernandes.

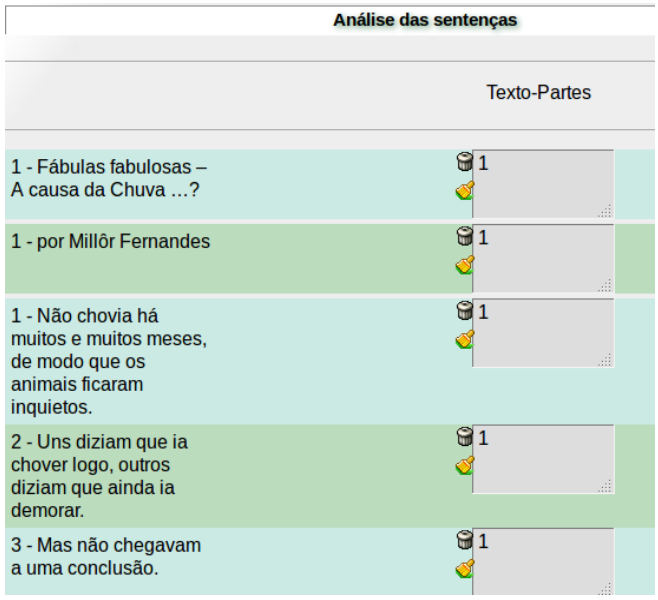

Ilustração 4: Sentenças iniciais de "A Causa da Chuva".

Note que as sentenças podem ou não ser do tamanho de um parágrafo, isso depende exclusivamente da análise morfossintática (Figura 4).

É possível editar as frases após o upload, mas é importante considerar que essa edição pode afetar a relação das mesmas com a análise morfossintática registrada durante o upload e que não pode ser editada NA VERSÃO 1.0 do dadosSemiotica.

a) Registro de textos

O conjunto de textos, cuja permissão de upload é exclusiva do gerente, deve estar no sistema antes da criação do projeto, sem impedir inclusão de novos testes depois. O nome do texto deve ser significativo para qualquer pessoa e deve seguir um padrão dentro da instalação do dadosSemiotica em uso, preferencialmente definida pelo gerente. Se você não é o gerente, converse com ele sobre o padrão a ser seguido.

NA VERSÃO 1.0, todos os analistas cadastrados numa instalação do dadosSemiotica tem acesso a todos os textos disponíveis no sistema. Se houver restrição de uso de alguns arquivos, em função de permissões junto a conselhos de ética ou direitos autorais, isso deve ser marcado no título do texto no sistema. Caso você não participe de um grupo com contato direto com o gerente e analistas, se seu texto tiver restrições desse tipo é preferível usar uma instalação particular. Isso não se aplica às análises, somente acessíveis pelo analista que as realizou.

O arquivo de entrada deve estar em txt, sem formatação, exceto por demarcação de parágrafos, e pode conter um texto completo. Nas versões 1.x recomenda-se, no entanto, dividir o texto em vários arquivos para evitar um longo tempo de pré-processamento (que causa demora no *upload* e pode incorrer em erro, devido a falhas na conexão). O ideal é não passar de 800kb.

O dadosSemiotica foi concebido para análise de texto verbal, portanto somente é possível associar análise de imagens e sons pela indicação de URLs no arquivo de texto de entrada (ou pela indicação de nomes de arquivos ou nomes de imagens ou sons que você possua localmente). Isso resulta, naturalmente, em perda de usabilidade, pois não será possível visualizar/ouvir o texto a ser analisado na mesma janela de análise. A inclusão de imagens no histórico do projeto é uma das funcionalidades previstas para a versão 2.0, o que permitiria mais facilmente incluir a visualização das mesmas em versões futuras.

A codificação do arquivo deve ser, preferencialmente, UTF8. NA VERSÃO 1.0 foram relatados problemas com arquivos vindos do Windows e gerados a partir de PDFs e DOCs. Os problemas foram resolvidos na versão 1.5 do sistema.

b) registro de categorias

A definição das categorias é essencial para a realização de um projeto no dadosSemiotica: com elas definimos todos os parâmetros de análise, inclusive o próprio escopo teórico.

NA VERSÃO 1.*, você deve criar as categorias da seguinte forma[4]:

i) a primeira palavra sempre é a categoria mais geral, correspondendo ao escopo teórico; se for uma categoria dentro de uma teoria, a primeira palavra deve remeter à teoria em si e deve ser iniciada por letra maiúscula. Se tiver mais de uma palavra, use maiúsculas em todo início de palavra e não deixe espaço entre elas. Por exemplo: GramáticaNormativa.

ii) a segunda palavra é a categoria de análise. Por exemplo, poderíamos querer analisar os vocativos, então a categoria seria GramáticaNormativa-vocativos. Sempre separando as duas por um traço sem espaços.

iii) podem haver tantas subdivisões nas categorias quantas forem necessárias, sempre seguindo a lógica da mais geral para a mais específica e separando-as por traços, sem espaços.

iv) é interessante registrar a categoria mais geral mesmo quando, num primeiro momento, não pareça fazer sentido analisá-la. Mesmo que ela não tenha uma análise específica, poderá ser usada para um primeira leitura do texto em relação ao nível de análise cujas subcategorias se referem, auxiliando nas análises mais específicas.

É possível cadastrar as categorias uma a uma ou entrar com todas de uma vez pela importação de um arquivo. A importação do arquivo CSV foi devidamente testada na versão 1.0 e o arquivo, que deve ser de texto simples, sem formatação, deve conter uma categoria por linha, sendo que, na primeira linha, deve conter apenas a palavra “nome”, sem aspas (figura 5).[5]

4 Esse método para a criação dos nomes das categorias visa facilitar o upgrade para a versão 2.0, que deverá trabalhar com grupos/hierarquias de categorias.

```
nome
Texto-partes
Texto-comentários-geral
Texto-comentários-lembretes
Semiótica
Semiótica-narrativa
Semiótica-narrativa-performance
```

Ilustração 5: Conteúdo de um arquivo para importação de categorias. O arquivo deve ser salvo num editor de texto, como texto sem formatação, em codificação UTF8, e ser salvo com a terminação .csv (exemplo: categorias.csv)

c) Categorias gerais - Texto

Algumas categorias são gerais, para o texto, independentes de contexto teórico. O grupo Texto contém essas categorias. NA VERSÃO 1.0, você deve registrá-las no sistema, se ainda não o fez. Sugerimos as seguintes categorias de análise para o grupo Texto:

Texto => esta categoria não deve ser criada: ela já existe no sistema e automaticamente é apresentada em todas as tabelas geradas, com o número do texto analisado como resultado. É especialmente importante quando o projeto contém mais de um texto-objeto.

Texto-partes => toda análise textual precisa de balizas, que podemos chamar de momentos, etapas, partes ou outros nomes. Podemos dividir o texto de qualquer maneira, seja por balizas definidas no próprio texto, como capítulos de um livro, seja por balizas que nossa leitura inicial indique (quando percebemos mudanças superficiais que podem indicar alteração no rumo das análises).

Texto-comentários => antes de iniciar uma análise textual, é sempre recomendável ler o texto inteiro. Como o dadosSemiotica divide o texto em sentenças, essa primeira leitura pode ajudar a perceber a relação entre as sentenças. Esta categoria serve para que você possa fazer anotações sobre essa leitura inicial e recuperá-las rapidamente em qualquer momento das análises posteriores. A natureza das anotações é totalmente livre.

Texto-lembrete => são anotações semelhantes às da categoria Texto-comentários, mas são mais específicas e de ordem prática: use essa categoria para anotar ações que deve ser feitas depois e que podem ser apagadas sem prejudicar a análise, tais como: “inserir uma nova divisão de parte aqui” ou “corrigir a ortografia dessa frase”. Depois de feito, basta apagar; o campo pode ter mais de um lembrete e você pode apagar parte dele quando o problema respectivo for resolvido. Você pode gerar uma tabela apenas com essa categoria para facilitar as correções depois, funcionando como um *check list*.

5 Com a criação de hierarquias de categorias e separação por escopo (teórico, principalmente), na versão 2.0 esse esquema de inclusão de listas de categorias poderia ser usado para a inclusão automática de uma lista em árvore.

5 Análise piloto para definir as categorias

O texto verbal é um sistema complexo e, mesmo no mais objetivo dos escopos teóricos, sempre sujeito a imprevisibilidades e quebra de regras. Por esse motivo, é recomendável que o analista, antes de iniciar as análises do seu projeto, realize uma pequena análise piloto com um trecho ou trechos dos textos do corpus de sua pesquisa, sob todas as categorias que pretende utilizar. Durante essa análise piloto é possível que se note a necessidade de uma maior especificidade nas categorias ou a não adequação de alguma(s) delas tendo em vista os objetivos do projeto em questão.

Essa análise piloto pode ser feita na forma de um novo projeto no dS. No entanto, como NA VERSÃO 1.* ainda não é possível copiar análises de um projeto para outro, é possível considerar que a análise piloto seja feita no mesmo projeto dS que será ampliado depois para a análise definitiva, caso se deseje aproveitar as análises do piloto no projeto principal.