Guia para anotação de correferência

Evandro Fonseca[1], Vinicius Sesti[1], Aline Vanin[2] and Renata Vieira[1]

evandro.fonseca@acad.pucrs.br, vinicius.sesti@acad.pucrs.br,
aline.vanin@ymail.com, renata.vieira@pucrs.br

[1]Pontifícia Universidade Católica do Rio Grande do Sul
[2]Universidade Federal de Ciências da Saúde de Porto Alegre

1 Introdução

Este documento contém instruções para anotação de cadeias de correferência em textos da língua portuguesa. O objetivo da tarefa consiste basicamente na correção das saídas geradas por um modelo de resolução de correferências previamente construído[1]. De forma a auxiliar na tarefa de anotação/correção, disponibilizamos o CorrefVisual. Um recurso que possibilita a correção dessas cadeias por intermédio de uma interface gráfica (Seção 4).

2 Cadeia de Correferência

A Resolução de correferências é uma tarefa que consiste em identificar as diversas formas que um mesma entidade nomeada ou sintagma nominal pode assumir em um determinado texto. Em outras palavras, esse processo consiste em identificar determinados termos e expressões que remetem a uma mesma menção.

Agrupando esses termos obtemos cadeias de correferência, como podemos visualizar na Figura 1. Na figura temos um texto com a seguinte cadeia em destaque:

- A ministra da Justiça do país

- Elisabeth Guigou

- a ministra

São expressões no texto que se referem a uma mesma entidade. Um texto geralmente possui várias desses grupos de menções. Na figura podemos ver ainda uma cadeia sobre uma norma da união européia e uma sobre o patenteamento de genes.

Uma relação de correferência pode aparecer de diversas formas. Como ilustramos nos exemplos a seguir:

[1] http://ontolp.inf.pucrs.br/corref/

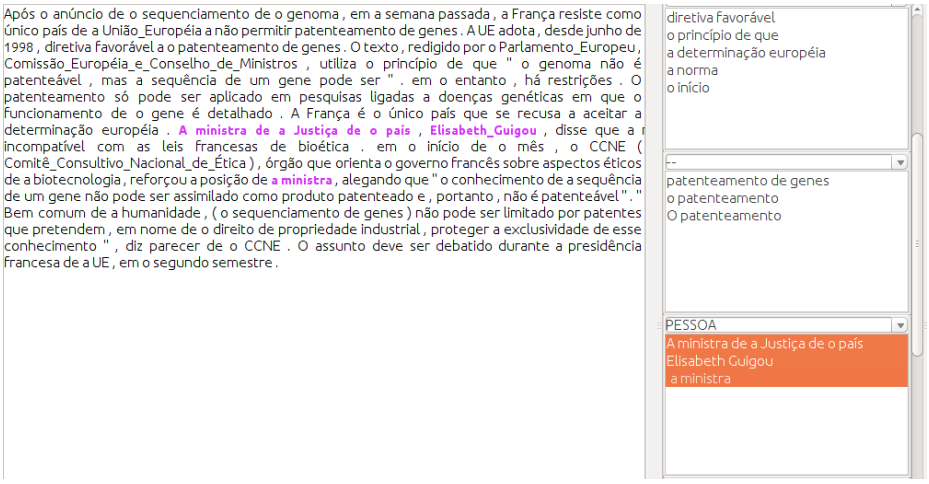

Figura 1. Cadeia selecionada em um texto

– O João está doente. Ele está com febre.

– A Ana comprou um cão. O animal já conhece todos os cantos da casa.

– Marcos está feliz. Seu gato passa bem.

– a pessoa era na verdade uma mulher de 20 anos.

– o menino, o garoto.

– Já se perguntou como as abelhas fabricam o mel? Os insetos saem em busca de...

– Cubatão, a cidade mais poluída do Brasil, localiza-se na Baixada Santista.

– O Instituto Nacional de Pesquisas Espaciais (INPE).

– Maria comprou várias frutas: mamão, melancia, abacate e uva.

Para facilitar a análise de menções correferenciais e não correferenciais, é possível analisarmos alguns padrões que podem indicar esse tipo de relação (subseção 3.1). É claro que essas não são as únicas formas em que uma relação de correferência pode aparecer. Contudo, são as mais frequentes. Na subseção 3.2 temos também alguns exemplos de menções não correferentes entre si.

2.1 Elementos da Cadeia

Referentes Como o próprio nome sugere, "referente" é a forma como nos referirmos a determinada entidade/sujeito. Em um texto, essas referências podem aparecer como uma entidade nomeada específica ou dentro de um sintagma nominal. Temos também correferência, que consiste na co-ocorrência dessas menções, referindo-se à mesma entidade/sujeito.

Entidades nomeadas São elementos utilizados para se fazer referência a objetos ou entidades de determinado discurso ou domínio. As classes podem ser nomes de pessoas, empresas, lugares, termos de alguma área específica, como genes, proteínas, entre outros. Por meio dos exemplos abaixo, podemos identificar diversas entidades nomeadas (ENs), como: Banco Nacional de Desenvolvimento Econômico e Social, Apple e bandas musicais.

- O Banco Nacional de Desenvolvimento Econômico e Social (BNDES), empresa pública federal, é hoje o principal instrumento de financiamento de longo prazo...
- A Apple informou que vendeu 5 milhões de iPhone 5 só em um fim de semana...
- Várias bandas de black metal tiveram influências do punk, tais como Venom, Celtic Frost, Bathory, Sarcófago, Darkthrone, Nazarene, Mayhem, Hellhammer, Behemoth, entre outras...

Sintagmas Nominais: São expressões linguísticas utilizadas para referenciar entidades em um discurso. No caso de um sintagma nominal, o núcleo pode configurar-se em nome comum, próprio ou um pronome. Os pronomes podem apresentar-se, basicamente, nas formas de pronome pessoal, demonstrativo, indefinido e possessivo. Note que um sintagma nominal pode também representar uma entidade nomeada, pois sua composição básica consiste de um determinante (artigo definido ou indefinido), seguido de um substantivo, o qual pode ser comum ou próprio. Contudo, um sintagma nominal pode ser composto por apenas um substantivo, seja ele comum ou próprio.

- a Microsoft;
- a empresa;
- o museu de Porto Alegre;
- museu;

- casa de cultura Mario Quintana;
- a região sul do Brasil;
- o Brasil.

2.2 Categorias das Cadeias

Além de serem agrupadas, em nosso modelo as cadeias são classificadas de acordo com as categorias descritas abaixo. Essas categorias são beaseadas no Repentino [2].

1. **Pessoa** – Representa nomes comuns, próprios ou profissões, que remetem à pessoas. Tais como: "Aline, Marcos, Barack Obama, menino, garoto, advogado, juíz, agrônomo, presidente...";
2. **Organização - Local** – inclui todas as organizações e locais, tais como nomes de empresas, de cidades, estados países. Nomes comuns também são levados em consideração, tais como: "praça, avenida, rua...";
3. **Eventos** – inclui-se todo e qualquer evento, cujo início ou duração estejam claramente definidos. Dentro desse contexto temos: "eventos esportivos, reuniões que envolvam qualquer atividade social ou cultural, como feiras e exposições; acontecimentos históricos, acontecimentos científicos (conferências, Simpósios);
4. **Comunicação** – incluem-se apenas entidades que são produtos relacionados com arte, mídia ou comunicação. Tais como filmes, livros, músicas, jogos digitais, publicações (como jornais e revistas); programas de tv, radio e teatro;
5. **Produtos** – inclui todo o tipo de produtos: comerciais, financeiros, farmacêuticos, industriais; tais como: ferramentas, eletrônicos, eletrodomésticos, produtos identificados por marcas (OMO, Aspirina-C, Clorofina...).
6. **Documentos** – incluem-se documentos em geral, tais como: leis, decretos, tratados, pactos, normas e planos.
7. **Abstração** – inclui entidades abstratas tais como disciplinas, ciências, processos (fotossíntese, pseudomorfose, sulfatação, osmose), teorias, doenças, estados condicionais, símbolos religiosos (crucifixo, pentagrama, Selo de Salomão...), crimes, índices de taxas (PIB, NASDAQ).
8. **Natureza** – incluem-se animais e vegetais assim como fenômenos naturais: ciclones, tufões, micro-organismos, elementos que constituem organismos vivos (músculos, células, ossos...).

9. **Outros Seres** – incluem-se todos os seres reais, ficcionais ou mitológicos, assim como os mitos, exceto pessoas e profissões (os quais são classificados como Pessoa). Exemplos: qualquer ser (real ou ficcional) que não seja humano, vivo ou morto; toda e qualquer entidade mitológica. Ex: Pégaso, Minotauro, Ícaro, Adamastor, Afrodite, Cupido, etc; grupos de pessoas (reais ou ficcionais) que partilhem a mesma identidade geográfica, política, étnica ou ideológica, embora não pertençam a uma organização estruturada, tais como: Incas, Budistas, Dadaístas, Nudistas, Marcianos, Atlantes, Visigodos, etc.

10. **Substâncias** – incluem-se elementos e substâncias como: Paracetamol, H2O, Anelina, penicilina, ácido ascórbico, acetilsalicilato de lisina, boldenona, hematoxilina, lecitina de soja, lidocaína, Mebendazol, nandrolona, Oxibutinina, álcool, glicose, etc.

11. **Outros** – Nesta categoria incluem-se exemplos que não foram encaixados em nenhuma das categorias utilizadas.

3 Menções correferentes e não correferentes

Nesta seção mostramos alguns exemplos de forma a facilitar a compreensão, referente as formas que podemos encontrar relações correferenciais entre duas ou mais menções. Sobretudo, é importante também tornar claro os casos em que temos falsos positivos. Ou seja, menções idênticas lexicalmente ou que possuem algum indício semântico de correferência, mas que remetem a entidades distintas. Para uma correta análise, é importante sempre levarmos em consideração o contexto de cada menção.

3.1 Menções correferentes

Comecemos com as formas mais comuns em que uma relação de correferência pode ser encontrada:

Menções idênticas: Geralmente em um texto podemos nos referir a uma mesma entidade a evocando da mesma forma que seu antecedente:

– Miguel Guerra participou do debate... Miguel Guerra relatou que...

Menções parcialmente semelhantes: Outra forma bastante comum de nos referirmos a uma mesma entidade é inicialmente a mencionarmos de forma específica e posteriormente, de forma mais genérica:

– Miguel Guerra participou do debate... para Guerra o país ainda...

Menções adjacentes:

- A Embrapa (Empresa Brasileira de Pesquisa Agropecuária);
- A ministra da justiça do país, Elisabete Guigou, informou que...
- o piloto Airton Senna.
- o animal é um clone gerado a partir de outro clone...
- Essas são as melhores férias que já tive.
- fazendeiros, cujas terras foram tomadas este ano.
- Um nome, o qual não pode ser revelado.

Siglas:

- a Organização das nações Unidas, a ONU;
- Pontifícia Universidade Católica do Rio Grande do Sul ... a PUCRS.

3.2 Menções não correferentes

Note que no exemplo abaixo temos duas menções falando de "o sul". Contudo, existem termos que as modificam ('da' e 'África'). Note que uma menção remete a "o sul do Brasil" e outra a "o sul da África". Portanto, não são correferentes.

- **o sul** do Brasil, **o sul** da África.

Menções que são parte constituinte de outra também não devem ser consideradas como correferentes entre si. No exemplo abaixo, note que a primeira menção fala de um automóvel com bancos de couro, já a segunda, fala de bancos de couro (entidades distintas).

- o carro com **bancos de couro**, **bancos de couro**.

Menções que possuem *Matching* exato, mas não remetem a um mesmo antecedente:

- o primeiro bebê$_1$ de 2008 nasceu em Lisboa às 00:00:30 segundos O provável primeiro bebê$_1$ português do ano é do sexo masculino e nasceu aos 30 segundos de hoje na Maternidade Alfredo da Costa, em Lisboa, disse à agência Lusa uma fonte daquele hospital. Com 3,520 quilos, o bebê$_1$ nasceu de parto normal e encontra-se bem, tal como a sua mãe, Mariana, de 16 anos.
  O provável segundo bebê$_2$ do ano nasceu no Hospital de São João, no Porto ao primeiro minuto do dia, de parto normal. Chama-se Francisco$_2$ e é o terceiro filho de um casal residente no Porto$_2$. O bebê$_2$...

Analisando o exemplo acima podemos notar que nem sempre um *matching* exato entre menções é o suficiente para torná-las correferentes. No exemplo citado temos referência a dois bebês (o primeiro e o segundo nascido). Note também que temos dois sintagmas "o bebê". Contudo, mesmo estes sintgmas sendo iguais lexicalmente, pertencem a cadeias distintas, pois remetem a entidades diferentes. Logo, para duas menções serem correferentes, não basta que sejam iguais, precisamos levar em consideração o contexto em que essas menções se encontram. Esse é um grande desafio, quando lidamos com esse fator em nível computacional.

4 CorrefVisual

O CorrefVisual é um recurso, desenvolvido com o propósito de auxiliar na correção de cadeias de correferência. A ferramenta apresenta uma interface amigável, permitindo editar as cadeias geradas automaticamente por nosso modelo[2]. Nesta Seção vemos o passo a passo para realizar as edições nas cadeias geradas automaticamente por nosso modelo. Basicamente, a utilização do CorrefVisual resume-se em quatro etapas:

Importação dos Dados – Basicamente, consiste em carregar o documento XML no programa (Figura 2).

Edição das cadeias – Esta etapa consiste na análise das cadeias de correferência geradas pelo modelo. Havendo uma ou mais menções posicionadas em cadeias incorretas, essas deverão ser corrigidas. Para executar tal ação, basta o usuário clicar e arrastar a menção para o grupo correto (*drag-and-drop*). Caso haja necessidade, é possível tambem criar novas cadeias, por meio do botão "novo grupo" (Figura 4). É importante tambem o usuário analisar a categoria semântica de cada cadeia de correferência. Caso seja necessário, é possível corrigi-la, aferindo uma nova categoria (ver Seção 2.2).

Edição de sintagmas – O usuário pode, além de editar que sintagmas pertencem a quais cadeias, também editar os próprios sintagmas. Isso é feito por meio dos quatro botões próximos ao campo "Editar sintagma", que adicionam ou removem um token no começo ou no final do sintagma.

Exportação dos Dados Modificados – Por fim, após completar as edições (se houverem), o usuário deve salvar suas edições. As subseções a seguir apresentam essas etapas de forma mais detalhada.

[2] http://ontolp.inf.pucrs.br/corref/

4.1 Importação dos Dados

Para importar um texto, basta selecionar o menu *Arquivo* > *Importar texto* e selecionar o arquivo XML correspondente. Feita a importação, é possível visualizar os três painéis principais do programa (Figura 2). O primeiro, à esquerda, responsável por exibir o texto puro; o segundo, central, contendo as cadeias de correferência; e o terceiro, à direita, contendo as menções únicas[3], juntamente com um painel auxiliar[4] e um painel de busca.

[3] Sintagmas que não pertencem a uma cadeia de correferência.
[4] Utilizado para o armazenamento temporário de menções, de forma a facilitar sua manipulação.

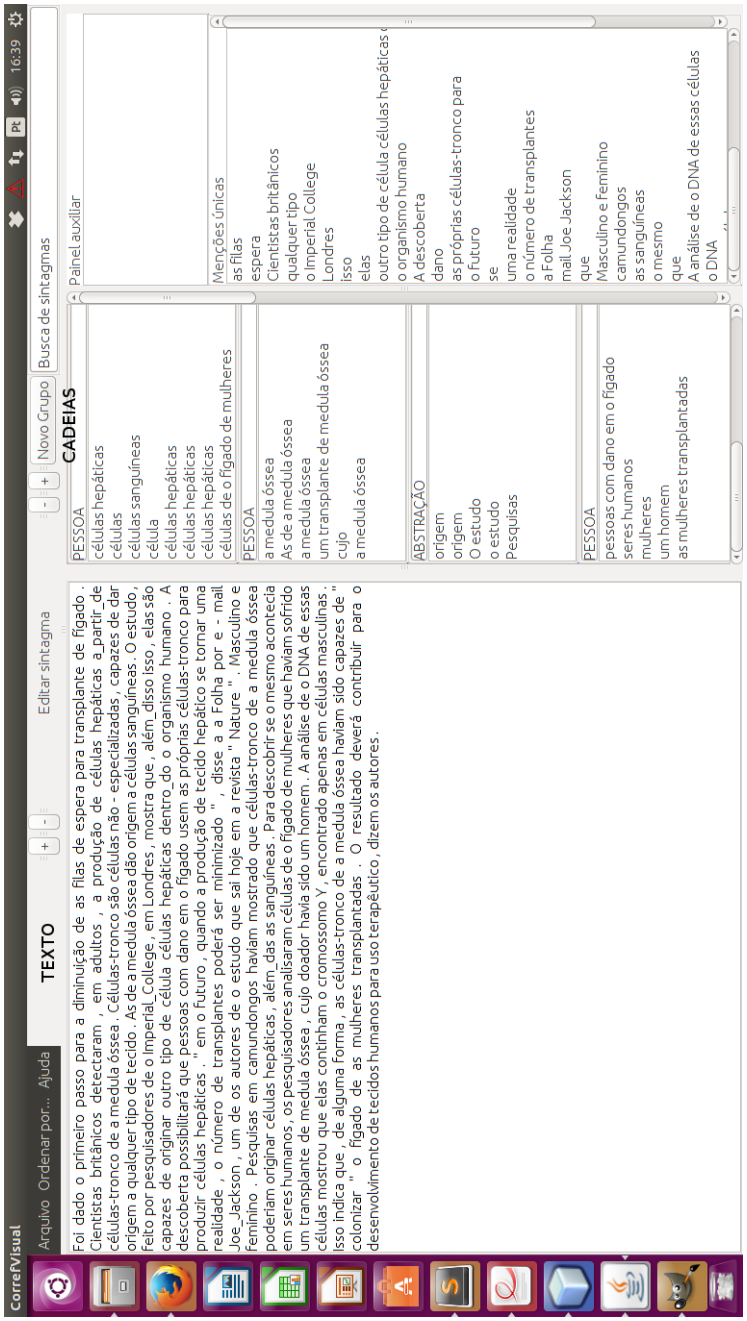

Figura 2. Tela principal do CorrefVisual

4.2 Edição das Cadeias e Categorias Semânticas

Nesta subseção serão mostradas algumas das funcionalidades que o CorrefVisual oferece em relação à manipulação das cadeias de correferência. A primeira delas é prover a visualização dos sintagmas e suas cadeias, facilitando a visualização de seu contexto. Basicamente, ao selecionarmos um ou mais sintagmas em uma cadeia ou do painel de menções únicas, eles serão coloridos no texto. Dessa forma, cada cadeia é marcada por uma cor única (por praticidade, todos as menções únicas partilham da mesma cor), conforme Figura 3

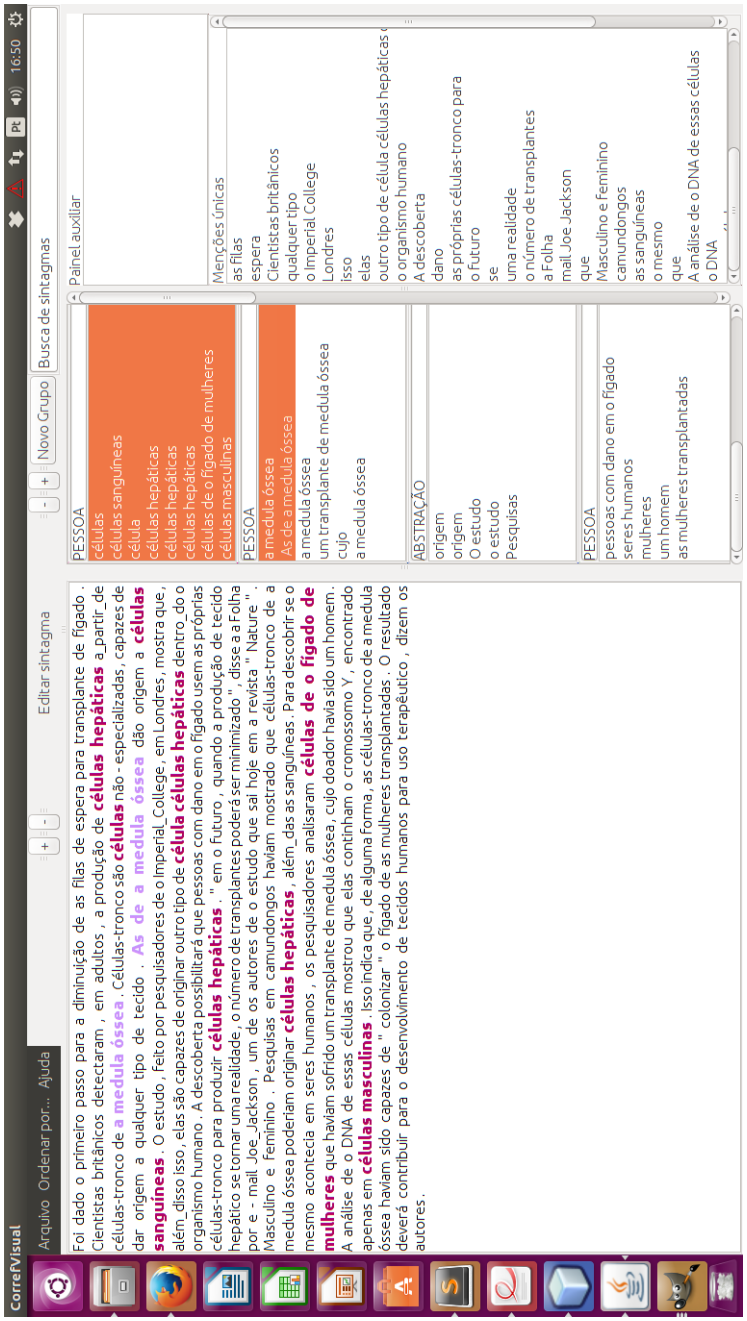

Figura 3. Seleção e coloração dos sintagmas nominais

É possível, também, arrastarmos sintagmas de uma cadeia para a outra conforme for necessário. Neste texto exemplo, os sintagmas "o conhecimento de a sequência de um gene" e "esse conhecimento" estão marcados como sendo menções únicas. No entanto, ao analisarmos a referência para este texto [1], percebemos que as duas menções pertencem a mesma cadeia. Para realizar a correção, criamos uma nova cadeia com o botão "Novo grupo" e arrastamos ambas as menções para este. O resultado pode ser observado na Figura 4.

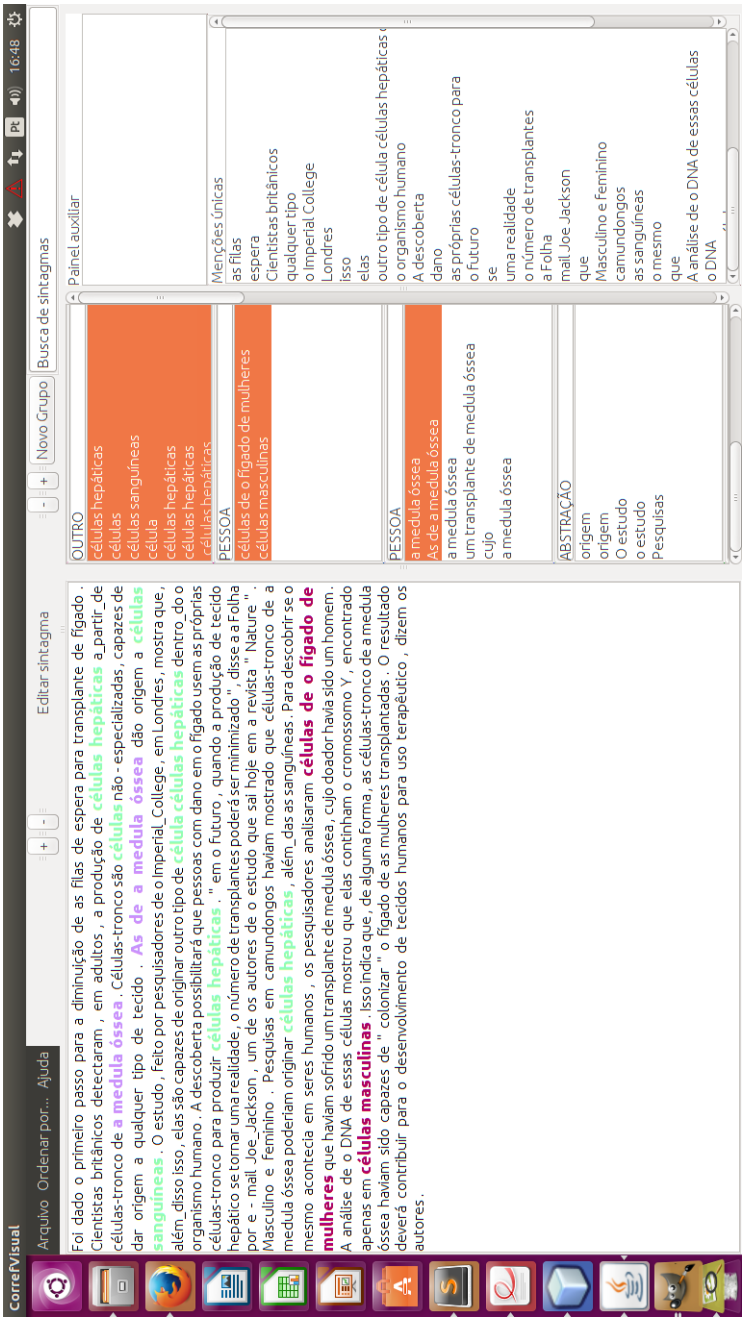

Figura 4. Criação e correção de uma cadeia de correferência

O CorrefVisual oferece a possibilidade de edição da categoria semântica das cadeias de correferência. As opções de categoria (ver Seção 2.2) aparecem como uma lista, acima de cada cadeia. Note que **todos** os sintagmas, dentro de uma mesma cadeia, têm a mesma categoria semântica, de forma que essa seleção vale para toda a cadeia.

Busca de sintagmas A funcionalidade de busca de sintagmas serve para buscar algum sintagma que contenha uma certa cadeia de caracteres ou palavra. Embora seja possível ordenar os sintagmas usando dois critérios[5] distintos, é possível que, ainda assim, seja difícil localizar algumas menções. Para buscar um sintagma, basta clicar na caixa "Busca de sintagmas", digitar o o sintagma desejado e apertar Enter. Em resposta, todos os sintagmas que contenham a cadeia de caracteres digitados, não diferenciando maiúsculas e minúsculas, serão selecionados.

Na Figura 5, a ferramenta foi usada para buscar sintagmas contendo o termo "célula". Podemos observar que todos os sintagmas contendo o termo foram selecionados.

[5] Ordem alfabética ou por ordem de aparição no texto.

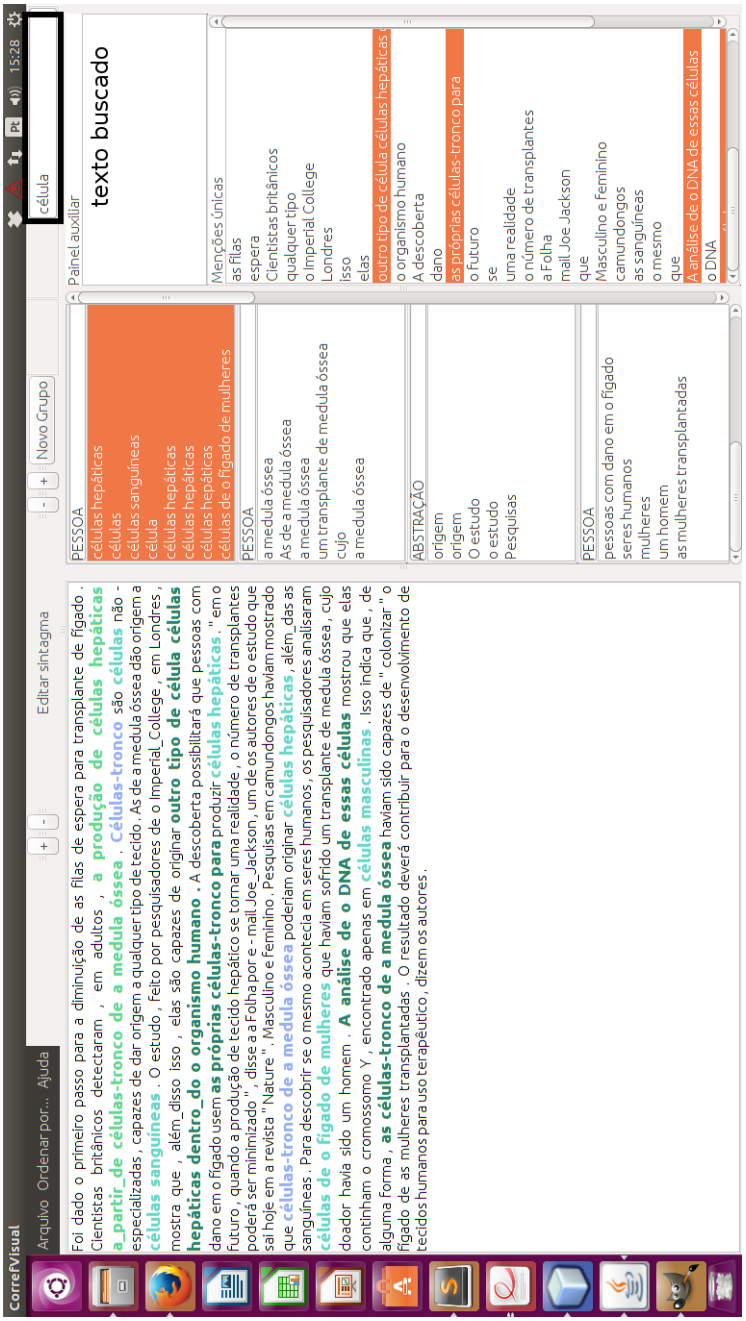

Figura 5. Busca por sintagma

4.3 Edição de Sintagmas

Outra funcionalidade oferecida no CorrefVisual é a edição de sintagmas, a qual provê uma forma rápida e eficaz, que permite alterar seus tokens. Assumimos que, no processo de chunking, o parser utilizado [3] é suscetível a erros e isso pode ocasionar em um truncamento incorreto dos adjuntos adnominais. Para permitir certa flexbilidade, é possível modificar os tokens pertencentes a cada sintagma por meio de quatro botões, localizados ao lado da caixa "Editar sintagma". Cada botão remove (botão -) ou adiciona (botão +) um token de um único sintagma.

Dentro desse contexto, os botões à esquerda da caixa adicionam/removem tokens no inicio do sintagma e; os botões à direita, adicionam/removem tokens no fim. Como exemplo, observa-se que temos selecionado um sintagma que era, anteriormente, "dano" e o editamos para "dano em o fígado", clicando três vezes no botão "+" direito. O resultado pode ser visto na Figura 6.

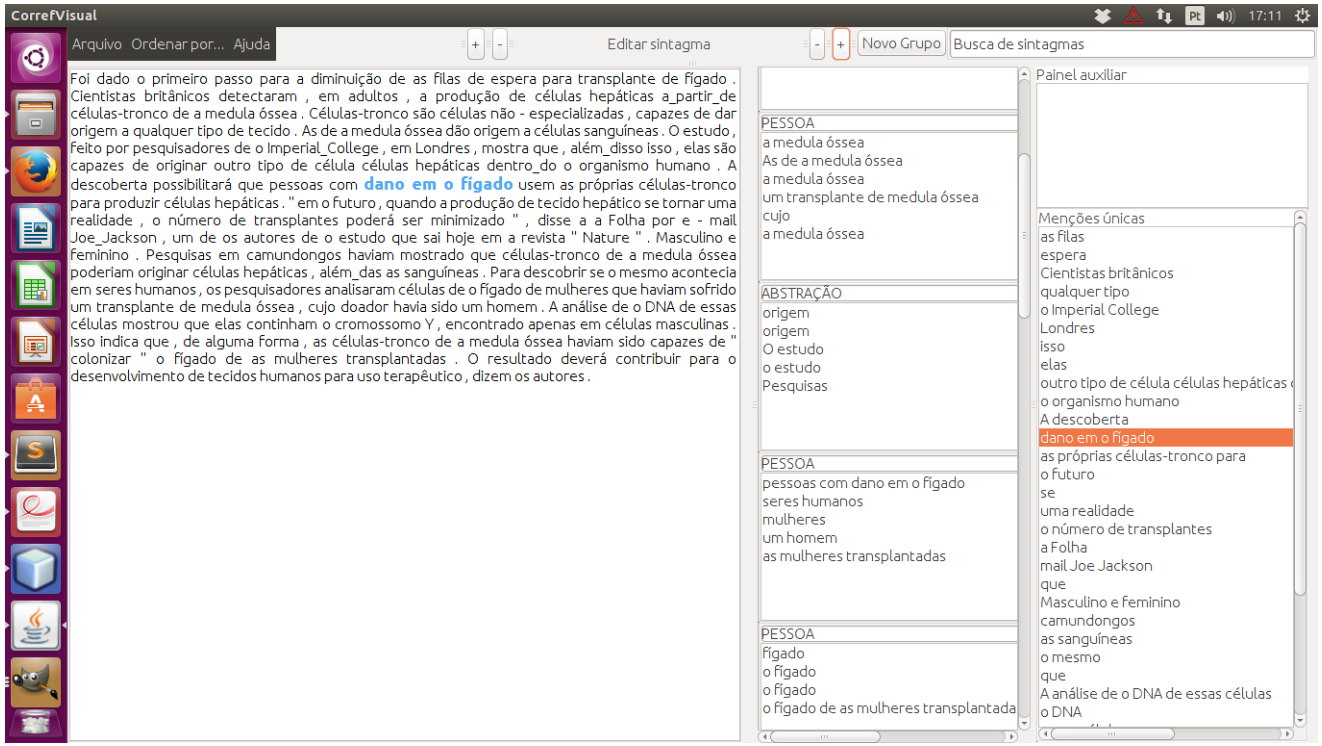

Figura 6. Edição de um sintagma

4.4 Exportação dos Dados

Para salvar as alterações feitas no texto, basta selecionar o menu *Arquivo* > *Salvar alterações*. Será criado um novo arquivo XML, estruturado da mesma forma, no diretório *saída*, presente na raiz do diretório onde o CorrefVisual está presente - este XML de saída tem o mesmo nome do arquivo de entrada. Conforme a Figura 7, uma mensagem é exibida na tela confirmando que o processo de salvamento foi executado corretamente.

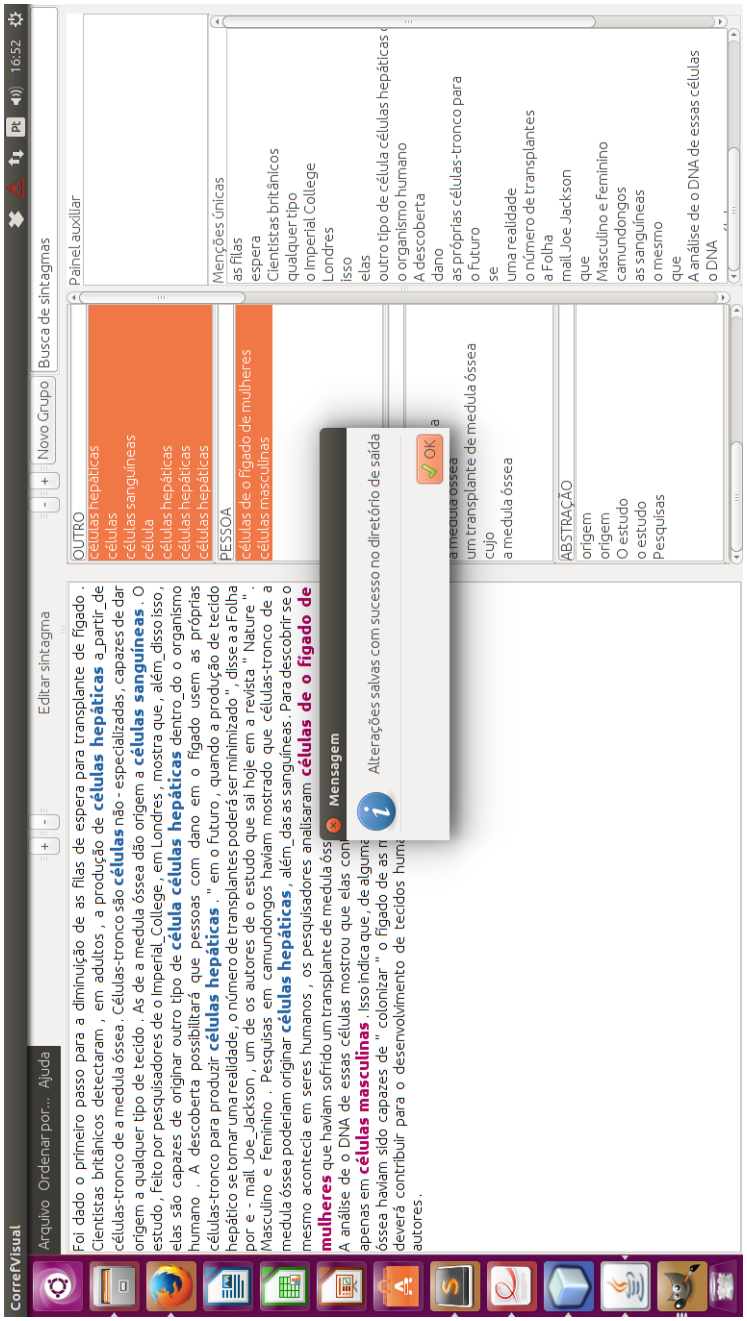

Figura 7. Exportação dos dados

Referências

1. A. Antonitsch, A. Figueira, D. Amaral, E. Fonseca, R. Vieira, and S. Collovini. Summ-it++: an enriched version of the summ-it corpus. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, In Press, 2016.
2. L. Sarmento, A. S. Pinto, and L. Cabral. REPENTINO - A wide-scope gazetteer for entity recognition in portuguese. In *Proceedings of Computational Processing of the Portuguese Language, 7th International Workshop - PROPOR, Itatiaia, Brazil*, pages 31–40, 2006.
3. W. D. C. Silva. Aprimorando o corretor gramatical cogroo. Dissertação de Mestrado, Universidade de São Paulo, 2013.