

NILC

Manual de uso NASP

Ferramenta de Anotação de Sentidos

Aluno: Fernando Antônio Asevedo Nóbrega
Orientador: Dr. *Thiago Alexandre Salgueiro Pardo*

Conteúdo

Introdução.....	1
Requisitos.....	1
Legenda da Aplicação.....	2
Descrição das telas.....	3
Processo de anotação	4
Marcações das anotações.....	4
Arquivos gerados durante a anotação.....	5
Exemplo de anotação.....	5
Exemplo de pré-anotação	7
Tecnologias utilizadas	8

Introdução

O NASP (Nilc – Anotador de Sentidos para o Português) foi desenvolvido para auxiliar o processo de anotação manual de sentidos (desambiguação lexical) de palavras do Português do Brasil usando a WordNet de Princeton (Wn-Pr) como repositório de significados.

Requisitos

1. Sistema operacional compatível com Java;
2. JVM (Máquina Virtual Java) instalada no sistema operacional (disponível em http://www.java.com/pt_BR/);
3. Conexão com a internet (desejável; caso não haja conexão com a Internet, algumas funcionalidades automáticas da ferramenta deverão ser executadas manualmente);
4. Os arquivos carregados para anotação devem ser codificados em UTF-8. Para verificar a codificação do arquivo, podem-se realizar as seguintes etapas (exemplificação descrita com softwares padrões dos sistemas Windows e Linux com idioma em português, caso o usuário esteja utilizando outro sistema ou idioma, devem ser executados comandos equivalentes):
 - a. Windows: abra o arquivo com o **bloco de notas** (notepad), clique em **arquivo** e depois em **salvar como**. Na janela que irá aparecer, é indicada a codificação do arquivo no campo **Codificação**, parte inferior da janela.
 - b. Linux (Ubuntu, distribuições Gnome): abra o arquivo com o **gedit**, clique em **arquivo** e depois em **salvar como**. Na janela que irá aparecer, a codificação do arquivo é indicada por **Codificação de caracteres**, na parte inferior da janela.

Caso o arquivo não esteja em UTF-8, deve-se selecionar essa codificação (nos campos apresentados para os respectivos sistemas operacionais) e salvar novamente o

arquivo.

Legenda da Aplicação

- a) Seletor de textos
- b) Visualizador do texto
- c) Lista de traduções
- d) Botão para adicionar uma nova tradução
- e) Opções para visualização dos *synsets* (hiperônimos e hipônimos)
- f) Lista de *synsets*
- g) Botão de seleção de *synset*
- h) Botão de remoção de anotação
- i) Lista de anotadores
- j) Botão para adicionar anotador
- k) Botão para editar o nome de um anotador
- l) Botão para remover um anotador

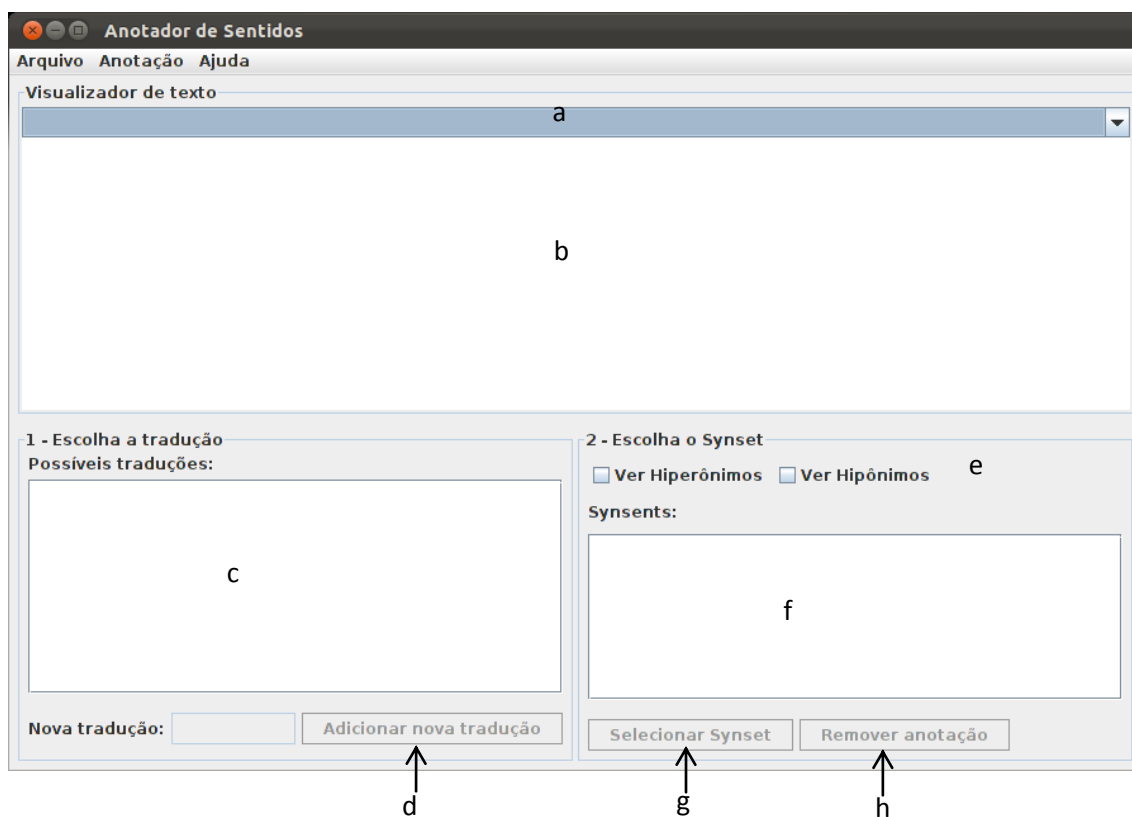


Figura 1:Tela principal

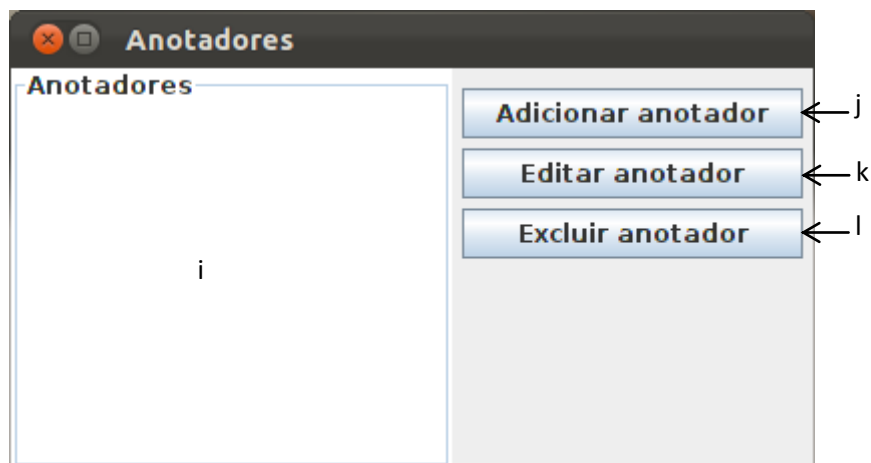


Figura 2: Tela de edição de anotadores

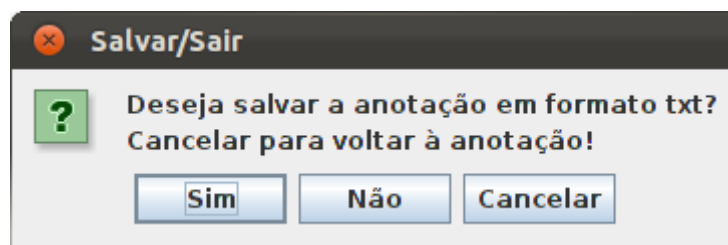


Figura 3: Mensagem de encerramento

Descrição das telas

A tela principal, apresentada na Figura 1, é destinada à anotação e também permite acessar outras funcionalidades do sistema, como: *abrir arquivos*, *abrir anotações anteriores* e *acessar a tela de edição de anotares*.

A tela de edição de anotadores, apresentada na Figura 2, permite cadastrar o nome dos anotadores que estão utilizando o software. Esse cadastro é opcional, porém, é indicado quando a tarefa está sendo realizado em grupo, permitindo manter um controle dos participantes. Essa tela é apresentada assim que o software é carregado e pode ser acessada por meio do menu **Anotação→Anotadores** (é possível adicionar mais de um anotador).

A mensagem de encerramento do sistema, apresentada na Figura 3, é apresentada ao usuário quando o software está sendo fechado. Esta tela permite ao usuário salvar a anotação em um formato texto (*TXT*) além do formato padrão *XML*, que é salvo automaticamente. Ao clicar em **cancelar**, o usuário volta para a anotação. Caso o usuário queira, também é possível salvar a anotação em formato texto sem encerrar o sistema, por meio do menu **Anotação→Salvar Anotação em Texto**.

Processo de anotação

O processo de anotação inicia após o carregamento de um texto, um grupo de textos ou um arquivo XML oriundo de anotações anteriores. Para cada palavra, a anotação ocorre em quatro etapas: (a) selecionar a palavra que será anotada; (b) selecionar/adicionar a tradução desta palavra para o inglês; (c) selecionar o *synset* correspondente; e (d) confirmar a anotação. Estas etapas são apresentadas na Figura 4.

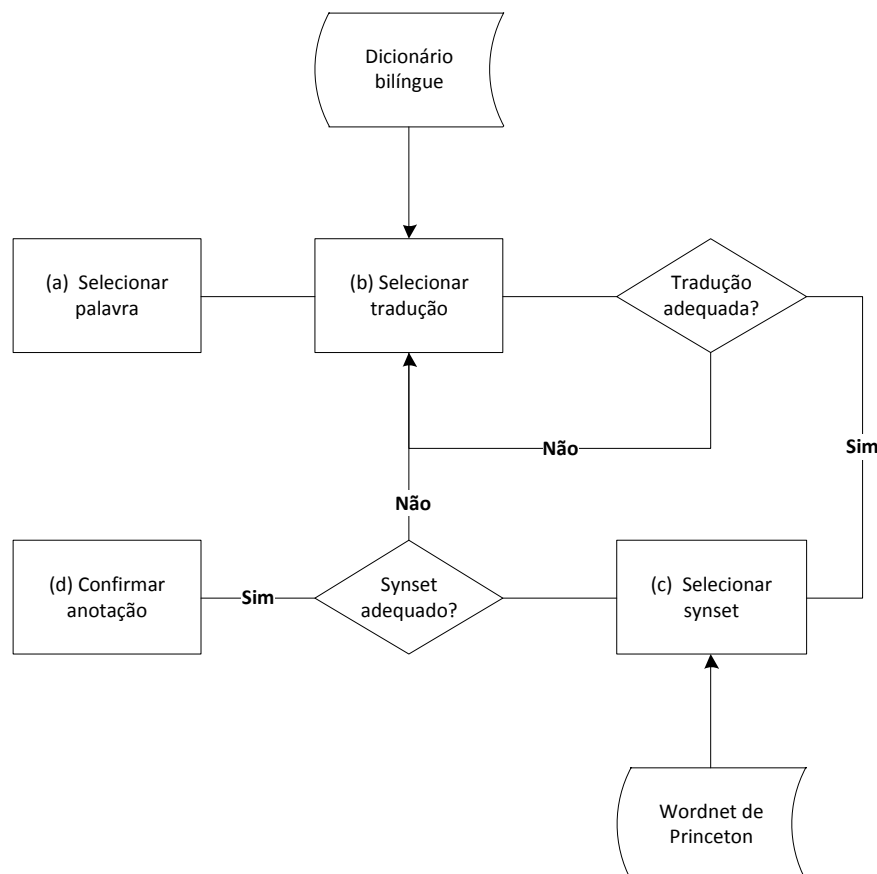


Figura 4: Diagrama do processo de anotação

É importante ressaltar que palavras com mesmo lema serão previamente anotadas após sua primeira ocorrência for desambiguada. Entretanto, o anotador pode alterar esta pré-anotação.

Marcações das anotações

Durante o processo de anotação, as palavras dos textos que estão sendo anotadas recebem marcações. Cada marcação, composta por uma cor e uma forma de visualização, representa um estado de anotação, como se apresenta a seguir:

1. **Fundo branco e borda vermelha**: substantivo comum ainda não anotado (ressalta-se que esses substantivos são identificados de forma automática, portanto, erros de etiquetagem podem ocorrer);
2. **Fundo e borda amarela**: palavra previamente anotada. Isso ocorre quando uma mesma palavra, com lema igual, foi anotada previamente; e
3. **Fundo branco e borda verde**: palavra anotada (desambiguada manualmente).

Arquivos gerados durante a anotação

Todos os arquivos (que não sejam arquivos temporários, usados pelo sistema) são gerados dentro do diretório **Texts**, que se encontra no diretório raiz da ferramenta. Quando um arquivo ou um conjunto desses é carregado para anotação, uma nova pasta é criada dentro do diretório Texts. Essa nova pasta irá representar o Diretório de Anotação (DA), que receberá todos os arquivos gerados durante o processo.

O DA é nomeado automaticamente pelo sistema, tendo o mesmo nome do diretório “pai” dos arquivos carregados. Por exemplo, se três arquivos de texto Texto_A, Texto_B e Texto_C são carregados, e todos estes arquivos estão no diretórioDiretorio_D, o sistema irá criar um diretório chamado Diretorio_D dentro do diretório Texts, onde serão salvos todos os arquivos gerados pela anotação.

São gerados três tipos de arquivos:

1. Anotação morfo sintática: demarcado pela extensão MXPOST
2. Anotação de sentidos em texto: um arquivo TXT demarcado por ANOTADO, que possui uma representação textual da anotação;
3. Anotação de sentidos: um arquivo XML que representa a anotação, assim no o formato no item 2, porém , possui informações extras utilizadas pela ferramenta de anotação. Esse arquivo é usado pela ferramenta para abrir anotações anteriores.

Exemplo de anotação

1. Carregando os textos a serem anotados. Na figura a seguir, é apresentada a tela de seleção de textos. Nessa tela, o usuário pode selecionar um ou mais textos (somente no formato txt e que se encontram em um mesmo diretório). Para acessar essa funcionalidade, deve-se acessar o menu **Arquivo →Abrir Arquivo(s)**. Outra possibilidade é abrir uma anotação salva anteriormente. Para isso, deve-se acessar o menu **Arquivo →Abrir Anotação** e selecionar um arquivo no formato xml. Quando mais de um texto é carregado, o anotador pode alternar a visualização entre os textos clicando na caixa de seleção do texto (item a apresentado na Figura 1).

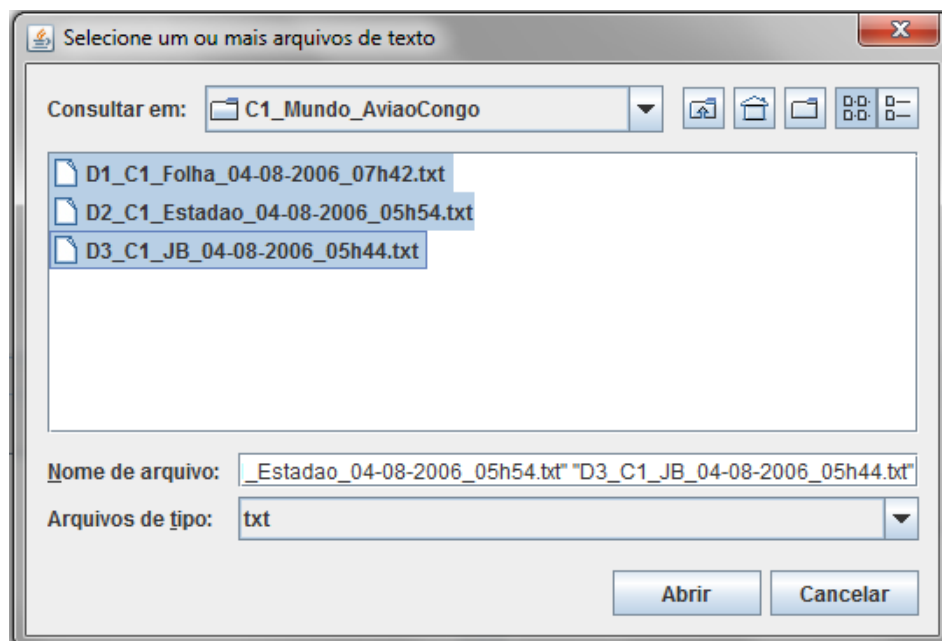


Figura 5: Tela de para selecionar os arquivos a serem anotados

2. A anotação ocorre por palavra, ou seja, anotam-se as palavras do texto uma a uma. Para selecionar uma palavra, basta clicar sobre a mesma. Por exemplo, a palavra “pessoas”, como apresenta na figura a seguir:

Ao menos 17 **pessoas** morreram após a **queda** de um **avião**

3. Selecionar uma tradução. Caso as traduções listadas automaticamente pelo sistema não sejam as melhores, o usuário pode inseri-las manualmente. Neste exemplo, é retornada apenas a tradução *person*.

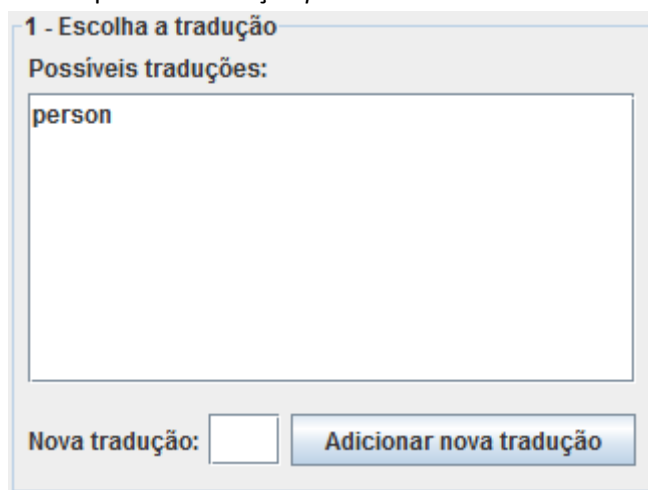


Figura 6: Lista de possíveis traduções

4. Selecionar o *synset* mais adequado. Após clicar em uma possível tradução, o software apresenta ao usuário todos os *synsets* ativados por essa tradução, como na figura a seguir. Por padrão, para cada *synset*, é apresentando o conjunto de sinônimos, o texto da glosa e todos os exemplos. Caso o usuário queira, também é possível visualizar todos os *synsets* que sejam hiperônimos e/ou hipônimos. Para

selecionar um synet basta clicar sobre o mesmo.

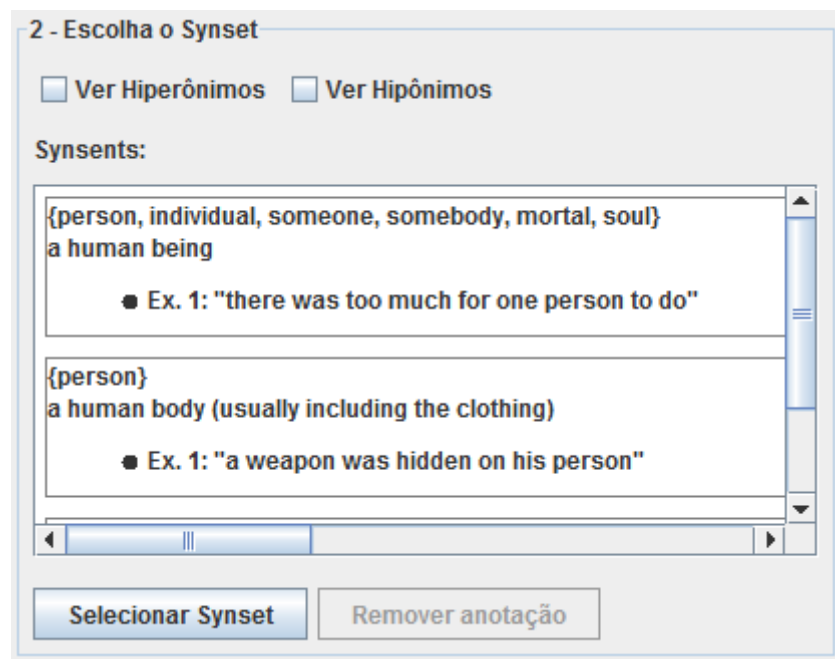


Figura 7: Lista de possíveis synsets

5. Anotar a palavra. Após selecionar a tradução e o *synset*, basta clicar no botão **Selecionar Sysnet**. Assim, a palavra será demarcada por uma borda verde e todas as demais ocorrências dessa palavra, que também não foram anotadas no texto, serão previamente anotadas.
6. Salvar anotação. O software salva o decorrer da anotação em formato XML a cada duas palavras anotadas e também ao encerrar a aplicação. Caso o usuário queira salvar a anotação, basta clicar no menu **Anotação**→**Salvar Anotação em Texto**.

Exemplo de pré-anotação

Uma palavra pré-anotada ocorre após uma palavra com mesmo lema ser anotada pela primeira vez no texto. Isso faz com que todas as demais ocorrências recebam a mesma tradução e o mesmo *synset* (podendo ser alterados posteriormente).

1. Clique sobre uma palavra pré-anotada (palavra marcada na cor amarela). Por exemplo, a palavra “avião” na figura a seguir

Segundo uma porta-voz da ONU , o avião , de fabricação russa

2. Após clicar em uma palavra pré-anotada, será carregada a lista de possíveis traduções (com a tradução selecionada, assim como apresentado na Figura 8, para a palavra *avião* do exemplo anterior) e a lista de *synsets* (também com o *synset* anotado selecionado, assim como apresentado na Figura 9) de sua primeira ocorrência anotada. O usuário pode aceitar esta anotação

simplesmente clicando no botão **Selecionar Synset** ou recomeçar a anotação desta palavra.

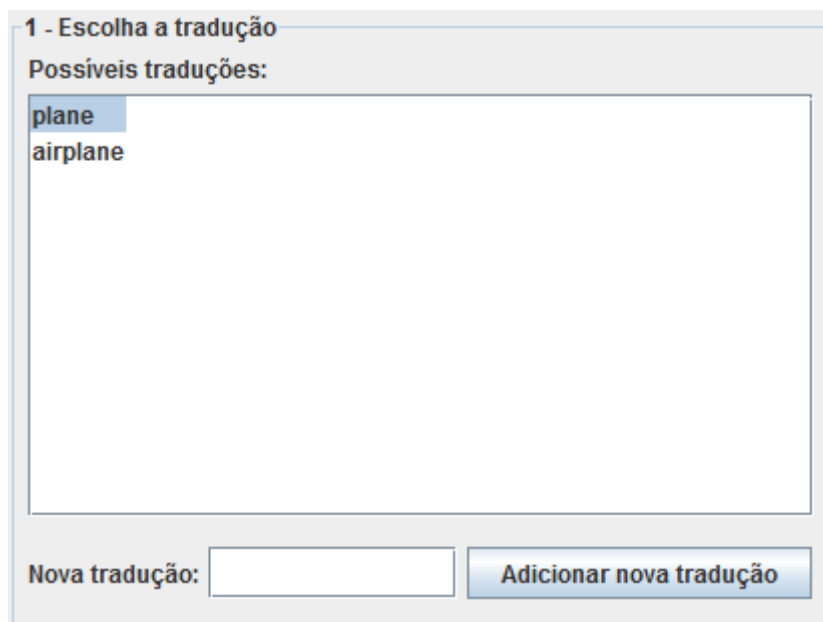


Figura 8: Exemplo de tradução selecionada por meio de uma palavra pré-anotada

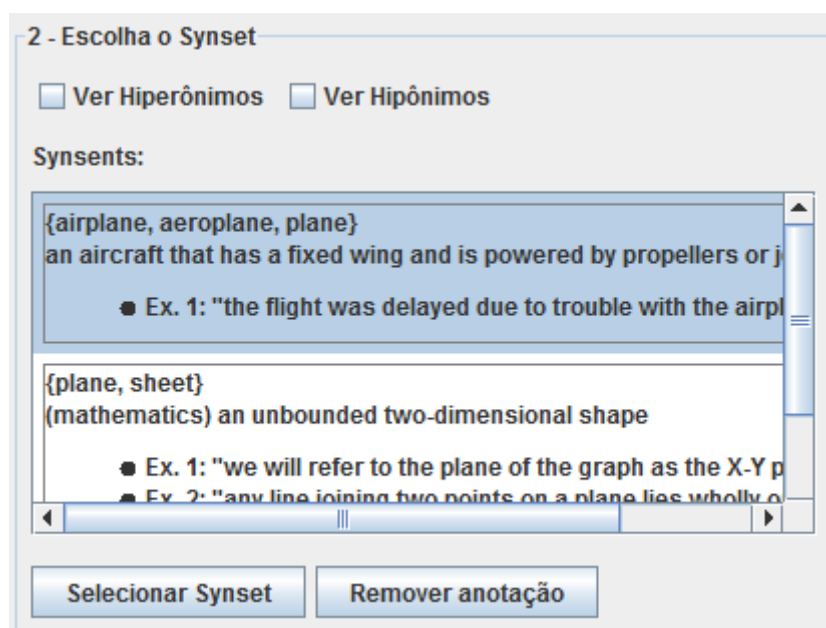


Figura 9: Exemplo de synset selecionado por meio de uma palavra pré-anotada

Tecnologias utilizadas

Este sistema foi desenvolvido usando a Linguagem de Programação Java, o que possibilita sua execução em diversos sistemas operacionais, desde que possuam a JVM instalada. De fato, essa tecnologia foi adotada para possibilitar a portabilidade do sistema. Entretanto, ressalta-se que a codificação dos arquivos carregados deve ser no padrão Unicode UTF-8.

Foram utilizadas três ferramentas computacionais: a base lexical da WordNet de Princeton, versão 3.0; o etiquetador morfossintático MXPOST (disponível em: <http://nilc.icmc.usp.br/nilc/tools/nilctaggers.html>); e o dicionário bilíngue WordReference (disponível online em <http://www.wordreference.com/>).

É Importante ressaltar que o etiquetador morfossintático MXPOST possui 97% de acurácia para o idioma Português. Essa métrica indica que há possibilidade de erros de etiquetação. Assim, a marcação prévia de palavras como substantivos (palavras com bordas em vermelho) é apenas um indicativo das palavras que podem ser anotadas, pois podem ocorrer duas situações: (1) Uma palavra foi etiquetada como substantivo, mas não é; e (2) uma palavra não foi anotada como substantivo, mas o é.