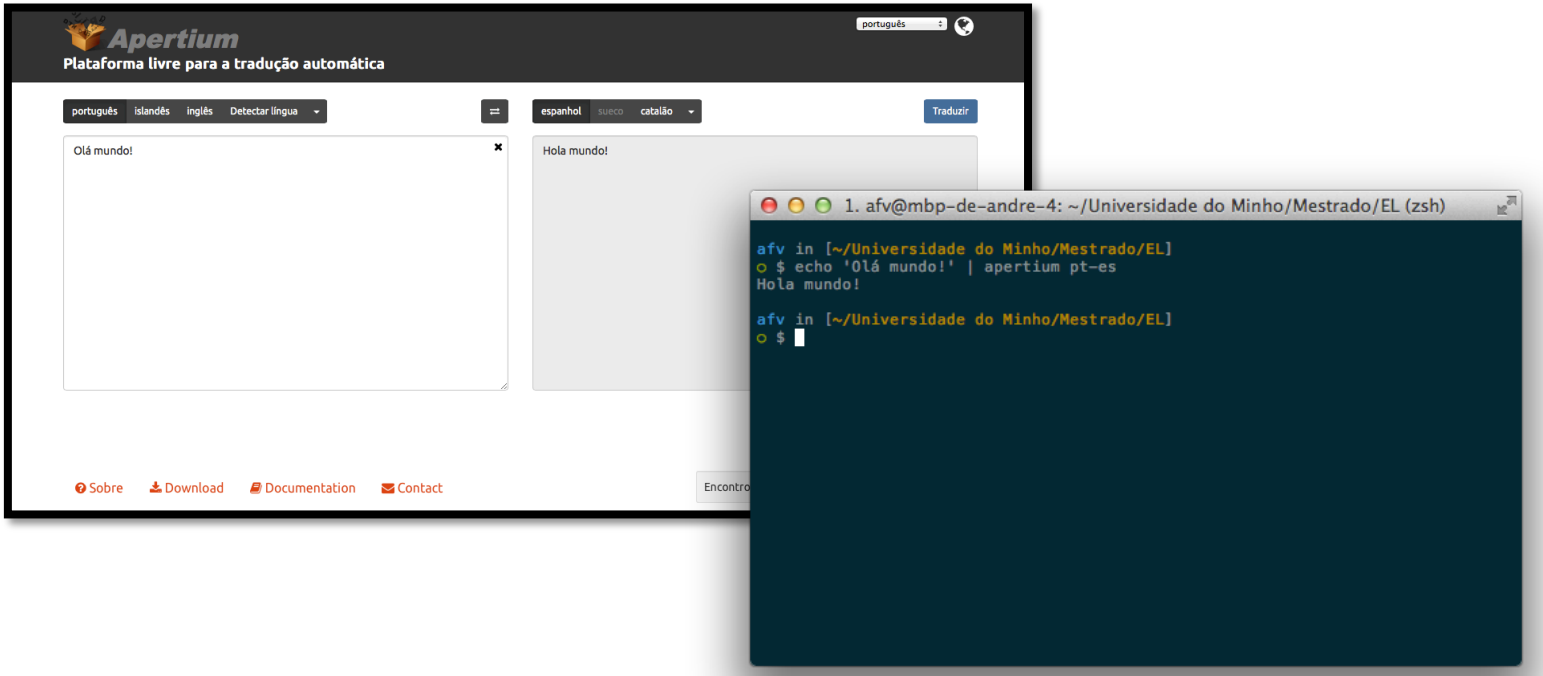
Apertium

A free/open-source machine translation plataform



André Vieira pg22777
Nuno Morais pg22806

Objetivos

- Apresentação do Apertium
- Instalação
- Termos
- Ficheiros base para tradução
- Descrição do seu funcionamento

Apresentação do Apertium

- O Apertium é uma plataforma de tradução que permite a construção de sistemas de tradução automática e que se originou com o projecto OpenTrad, o qual foi financiado pelo governo espanhol. Este projecto foi originalmente projectado com vista à tradução de línguas próximas, embora recentemente tenho sido expandido de modo a suportar pares de línguas mais divergentes;

- Para criar um sistema de tradução novo, é necessário desenvolver os dados linguisticos (dicionários, regras) especificados em formato XML;

- A plataforma foi criada em C++ e todos os ficheiros de dados estão em formato XML.

Apresentação do Apertium

- Um motor de tradução de uma linguagem em Apertium passa por 4 fases:
  - **Incubator** - É um local onde as pessoas colocam dicionários e outras coisas que possam ser úteis na construção de pares de línguas.
  - **Nursery** – É o local para onde os pares de línguas vão e que podem ser construídos, mas que não estão desenvolvidos completamente.
  - **Staging** – É para onde os pares de línguas vão quando estes estão desenvolvidos completamente, mas ainda não estão completamente preparados para lançamento. Pode faltar pequenos pedaços de código, ou algumas partes podem ter de ser ajustadas.
  - **Trunk** - É onde os pares de línguas são lançados. Eles devem ser funcionais.

https://svn.code.sf.net/p/apertium/svn/

Instalação

- **Macports**
  - sudo port install autoconf automake expat flex gettext gperf help2man libiconv libtool libxml2 libxslt m4 ncurses ncursesw p5-locale-gettext pcre perl5.8 pkgconfig zlib gawk subversion
- **Outros**
  - Ltoolbox;
  - Apertium;
  - Ficheiro de linguagem.

http://wiki.apertium.org/wiki/Installation

Ltoolbox

- **lttoolbox** é um conjunto de ferramentas para processamento lexical, análise morfológica e geração de palavras. *Esta análise* é o processo de separar uma palavra no seu lema e na sua informação gramatical. *Geração* é o processo oposto.

- **Cães**

  - **Lema** -> Cão

  - **Informação gramatical** - <n><m><pl>

Termos

- **Lema –** palavra sem flexão.

- **Símbolo/Tags** - etiqueta gramatical;

  - **<sg>** singular
  - **<pl>** plural
  - **<p1>** primeira pessoa
  - **<pri>** presente do indicativo

- **Paradigma** - Quando duas palavras são flexionadas da mesma maneira, criam-se regras que dizem que a palavra X é flexionada como a palavra Y.

Apertium - Base

- **São necessários três dicionários principais (formato dix):**
  - **O dicionário morfológico para a língua X**: contendo as regras relativas à flexão das palavras na língua X;
  - **O dicionário morfológico para a língua Y**: contendo as regras relativas à flexão das palavras na língua Y;
  - **Dicionário Bilingue**: contendo a correspondência entre palavras e símbolos nas duas línguas.

Pares de linguagens

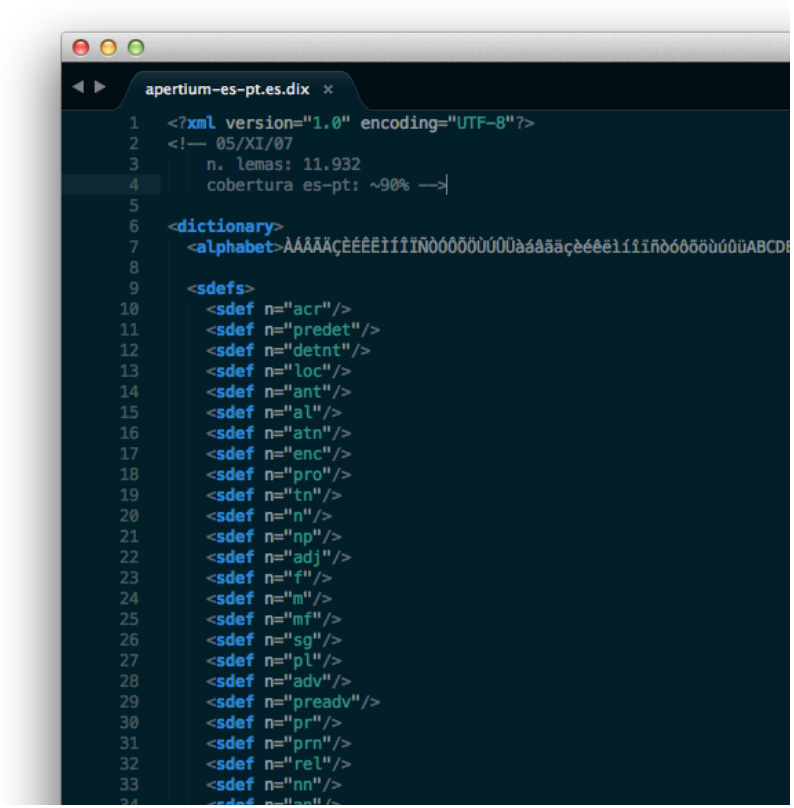
Translators

The following 37 pairs have released versions and are considered to be stable:

- Spanish ⇆ Catalan (es-ca)
- Spanish ← Romanian (es-ro)
- French ⇆ Catalan (fr-ca)
- Occitan ⇆ Catalan (oc-ca)
- English ⇆ Galician (en-gl)
- Swedish → Danish (sv-da)
- Macedonian → English (mk-en)
- Afrikaans ⇆ Dutch (af-nl)
- Indonesian ⇆ Malaysian (id-ms)
- Icelandic ⇆ Swedish (is-sv)
- Occitan ⇆ Spanish (oc-es)
- Spanish ⇆ Portuguese (es-pt)
- English ⇆ Catalan (en-ca)
- English ⇆ Spanish (en-es)
- English ⇆ Esperanto (en-eo)
- Spanish ⇆ Asturian (es-ast)
- Catalan ← Italian (ca-it)
- Maltese → Arabic (mt-ar)
- Serbo-Croatian ⇆ Slovenian (hbs-slv)
- Spanish ⇆ Galician (es-gl)
- French ⇆ Spanish (fr-es)
- Esperanto ← Spanish (eo-es)
- Welsh → English (cy-en)
- Breton → French (br-fr)
- Icelandic → English (is-en)
- Esperanto ← Catalan (eo-ca)
- North Sámi → Norwegian (sme-nob)
- Serbo-Croatian → Macedonian (sh-mk)
- Portuguese ⇆ Catalan (pt-ca)
- Portuguese ⇆ Galician (pt-gl)
- Basque → Spanish (eu-es)
- Norwegian Nynorsk ⇆ Bokmål (nn-nb)
- Macedonian ⇆ Bulgarian (mk-bg)
- Esperanto ← French (eo-fr)
- Basque → English (eu-en)
- Spanish ⇆ Aragonese (es-an)
- Kazakh ⇆ Tatar (kaz-tat)

Dicionários

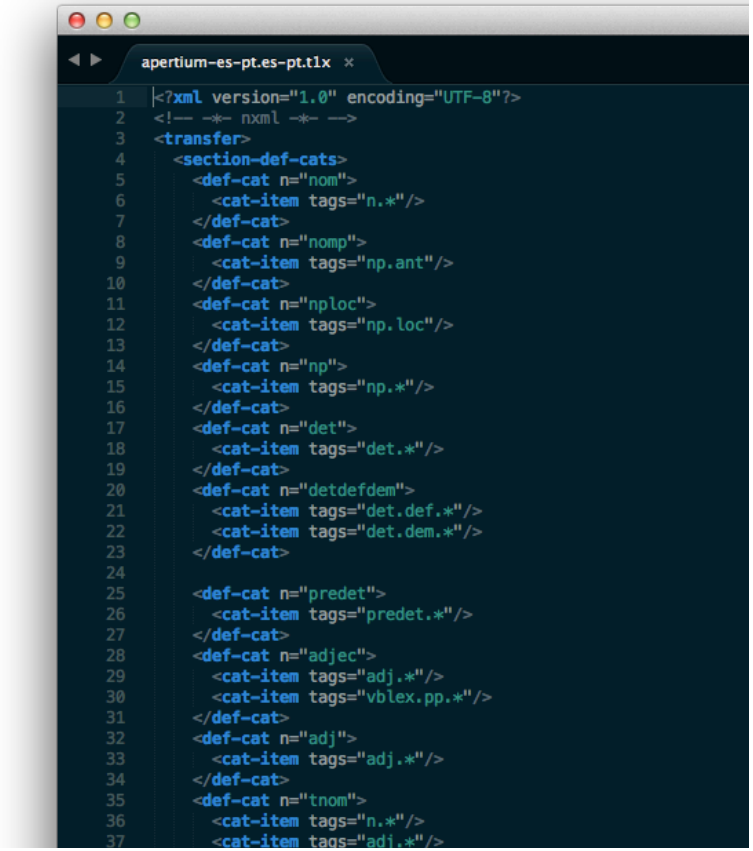
- Os ficheiros de dicionário são ficheiros XML onde é especificado o alfabeto utilizado, são definidos os símbolos/ tags e os paradigmas;
- Está dividido em duas secções: a secção standard que contem palavras e a secção incondicional que contém tipicamente a pontuação;
- O dicionário bilingue terá informação relativa aos pares da palavra nas duas línguas.



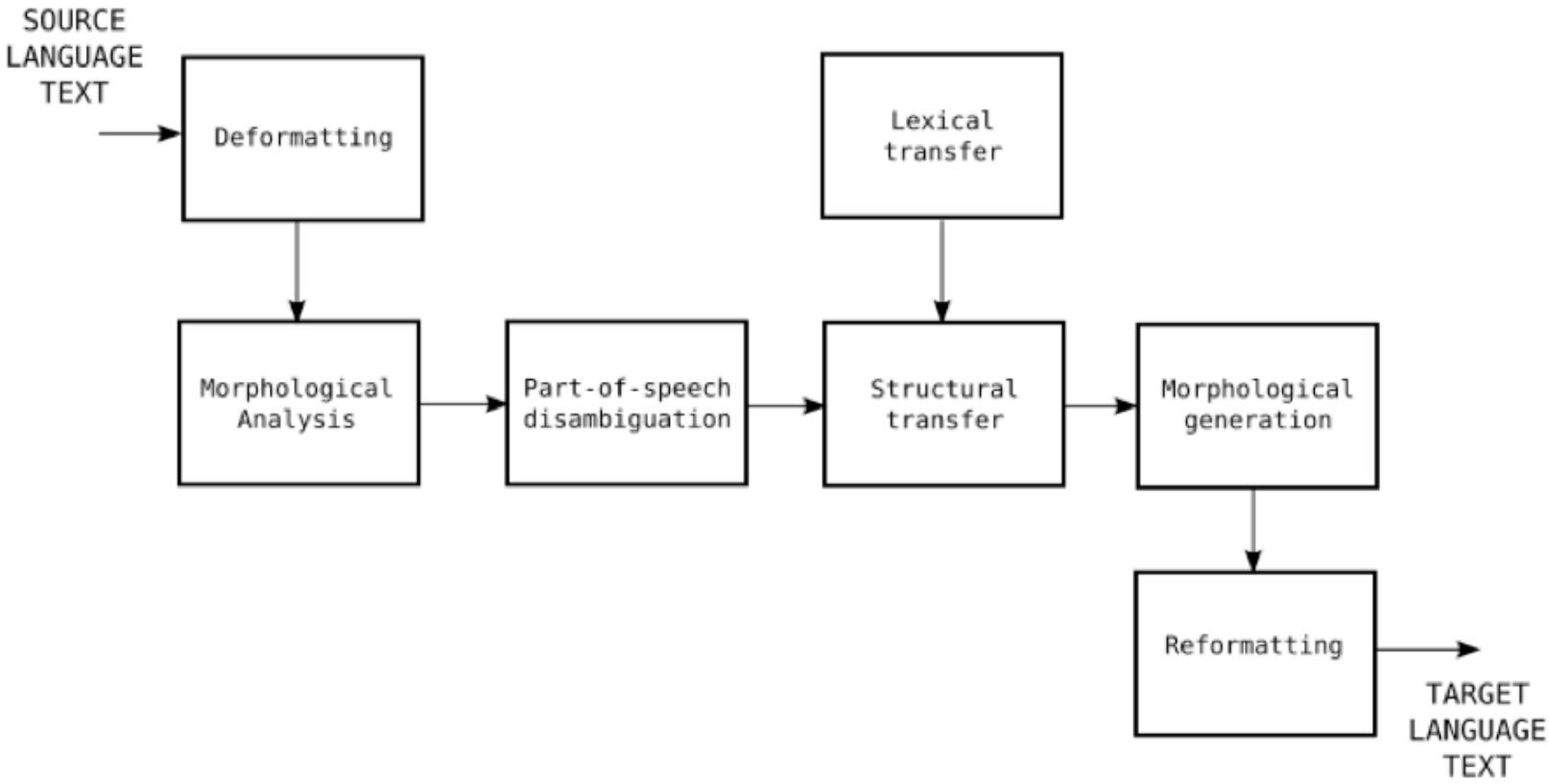
Ficheiros de regras

- Os ficheiros de regras de transferência são também ficheiros XML e descrevem uma série de acções quando um padrão é encontrado.

- Por exemplo, sendo encontrado o padrão <verbo> seguido do <tempo verbal>, é realizado uma determinada acção, que será conjugá-lo de maneira correcta consoante o tempo verbal descrito.



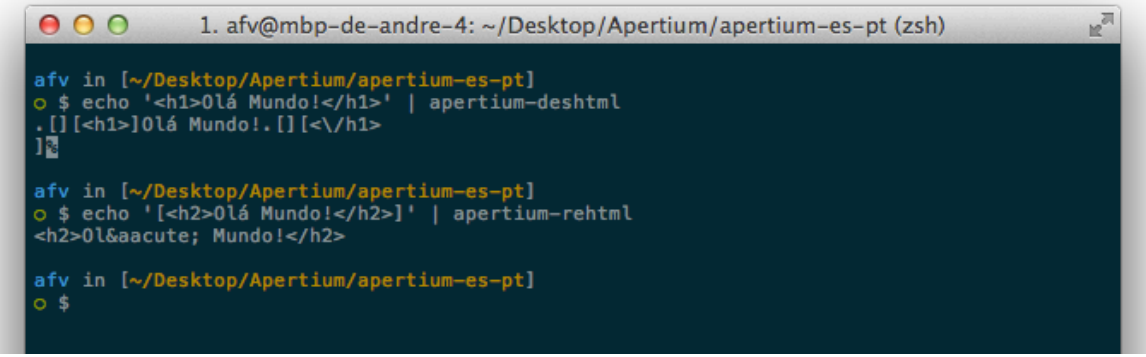
Como funciona?



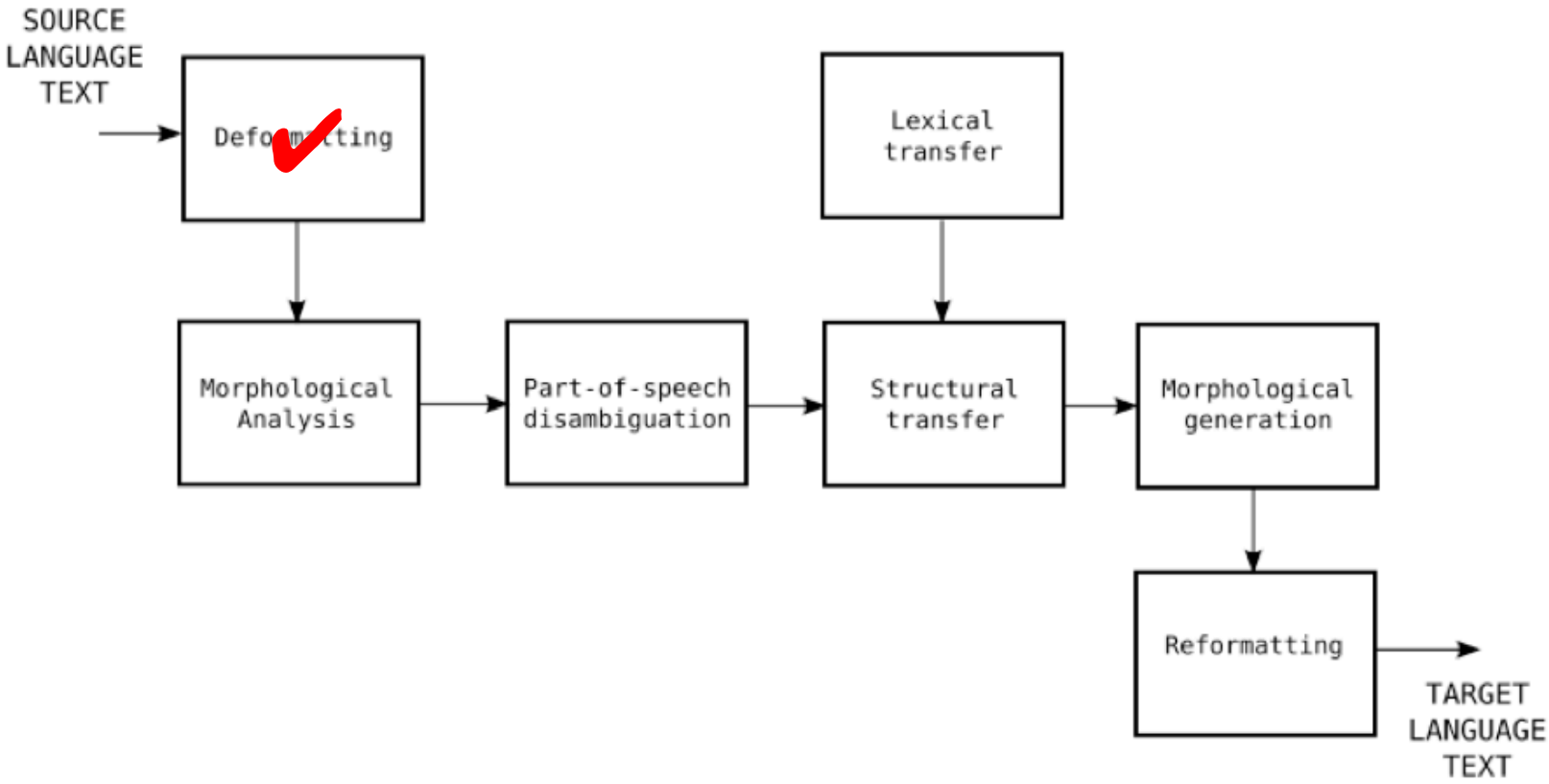
Deformatting

- Neste passo, a formatação de determinado ficheiro é encapsulada e delimitada pelos caracteres [e];

- Como exemplo, no caso do processamento de um ficheiro HTML, o programa apertium-deshtml encapsula a informação de formatação, enquanto que o apertium-rehtml restora essa informação.



Como funciona?



Morphological Analysis

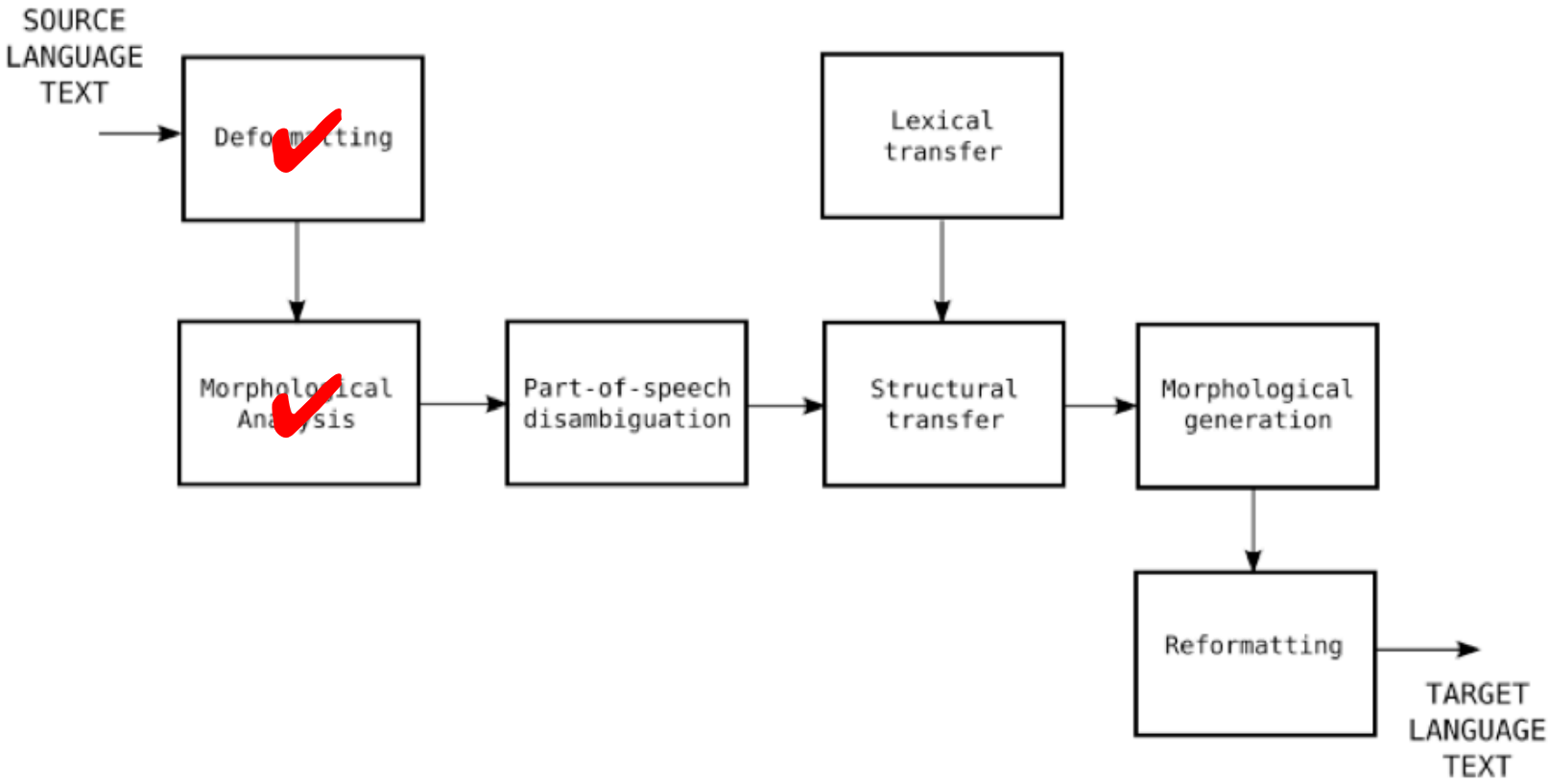
- A **Análise Morfológica** tenta modelar as regras que governam a estrutura interna das palavras de uma língua. Estas regras reflectem padrões específicos na forma como as palavras são formadas a partir de unidades menores e como essas unidades menores interagem entre si.

- **Exemplo**:

  - **Cão** -> Cães, Cãozinho;

    - **Lema**: Cão;
    - **Informação gramatical**: <n>, <m>, <pl>, <sg>, etc...

Como funciona?



Part-of-speech Tagging

- **Part-of-speech tagging** é o processo de atribuição de categorias gramaticais não ambíguas, para palavras num determinado contexto.

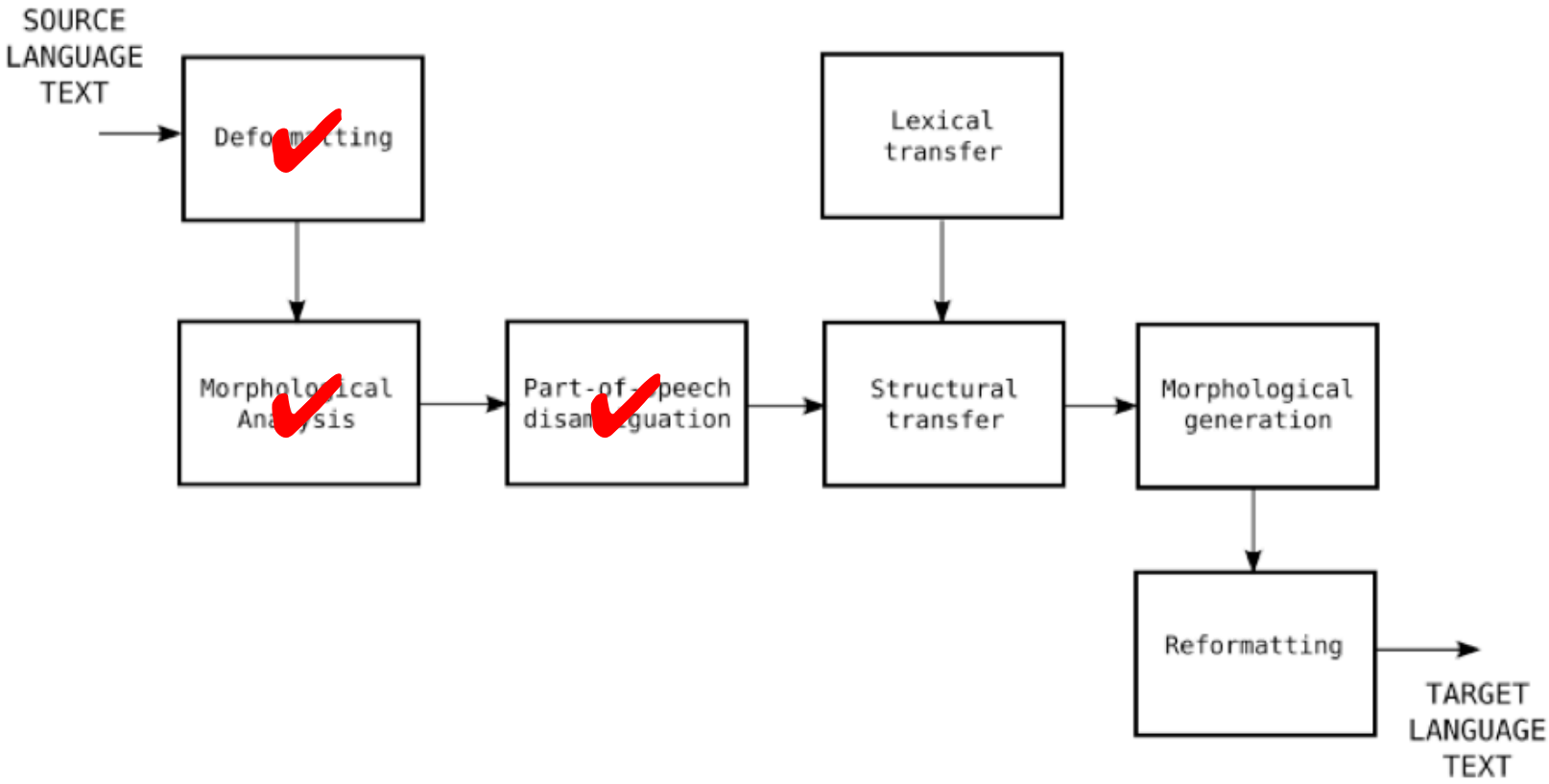
- O appertium-tagger, responsável por este passo, utiliza um modelo probabilistico (hidden Markov), que é “treinado” estatisticamente mas pode também ser influenciado por regras. Este calcula o caminho mais provável por entre uma sequencia de probabilidades.

Part-of-speech Tagging

Matrix of transition counts							
	Second tag						
First tag	verb	noun	det	prn	pr	adj	sent
verb	0	1	1	0	2	1	0
noun	1	0	0	0	1	1	3
det	0	4	0	0	0	0	0
prn	0	0	0	0	0	0	0
pr	0	1	2	0	0	0	0
adj	0	0	0	0	0	0	2
sent	3	0	1	0	0	0	0

Matrix of transition probabilities (A)							
	Second tag (γ_j)						
First tag (γ_i)	verb	noun	det	prn	pr	adj	sent
verb	0	0.2	0.2	0	0.4	0.2	0
noun	0.16	0	0	0	0.16	0.16	0.5
det	0	1	0	0	0	0	0
prn	0	0	0	0	0	0	0
pr	0	0.3	0.6	0	0	0	0
adj	0	0	0	0	0	0	1
sent[3]	0.75	0	0.25	0	0	0	0

Como funciona?



Lexical Transfer

- O modelo mais simples de tradução é a substituição de cada palavra pela sua correspondente no dicionário bilingue;

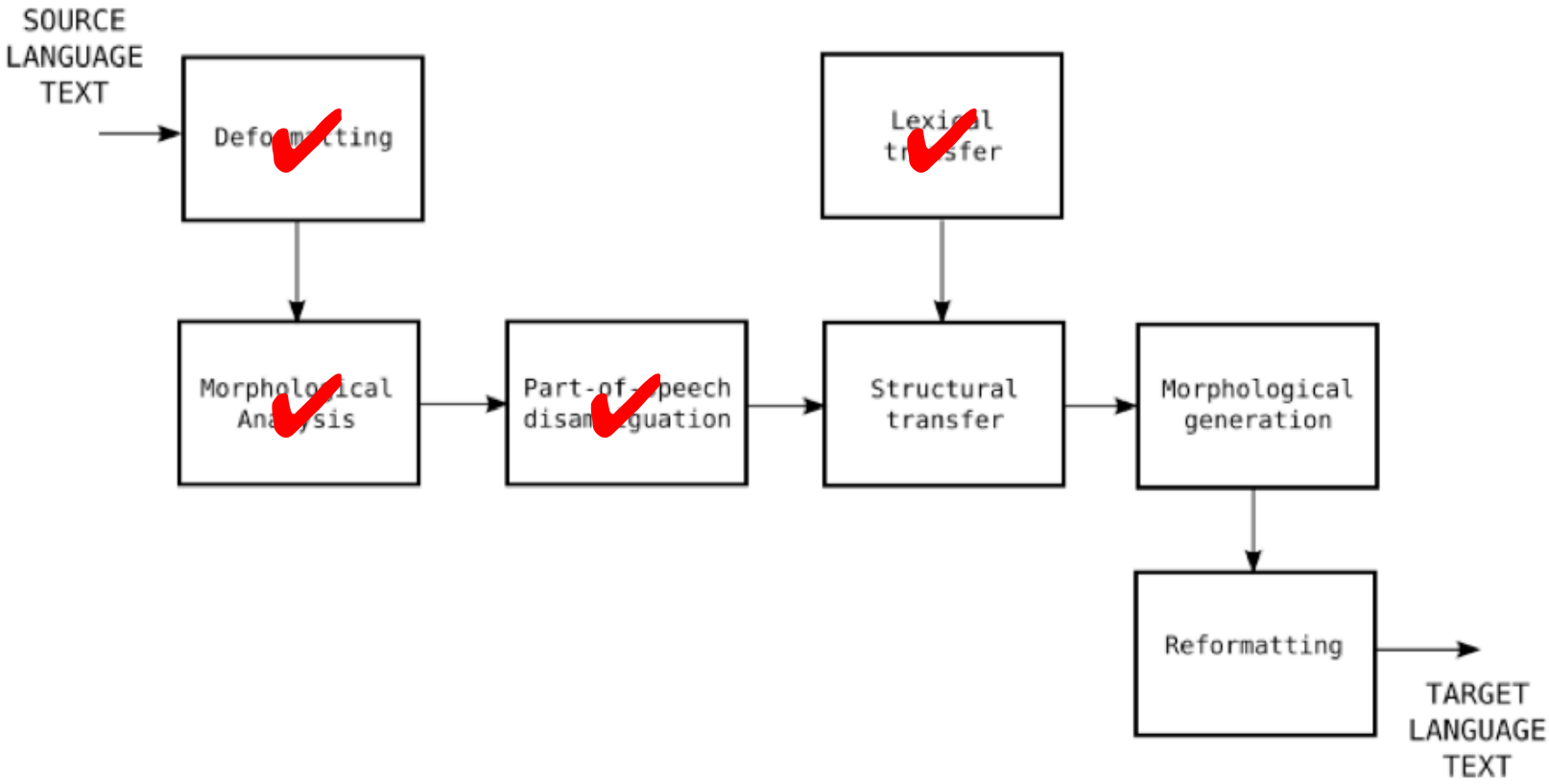
- Quando o apertium é utilizado para traduzir línguas menos relacionadas entre si, a questão de selecção léxica torna-se significativa pois existem casos onde uma palavra da língua de origem podem ter mais do que uma tradução na língua alvo. O módulo de selecção léxical lida com este problema.

- O problema é resolvido de duas maneiras:

  - No caso de ambas as traduções serem sinónimas e o sentido da frase não se alterar, é escolhido normalmente o lema de tradução mais usual;

  - No caso do significado das traduções ser diferente, é escolhida a correcta pelo meio de métodos estatísticos.

Como funciona?



Structural Transfer

- Neste passo dá-se o processo de alteração da estrutura representativa de uma línguagem de origem noutra referente à linguagem alvo seguindo um conjunto de regras;

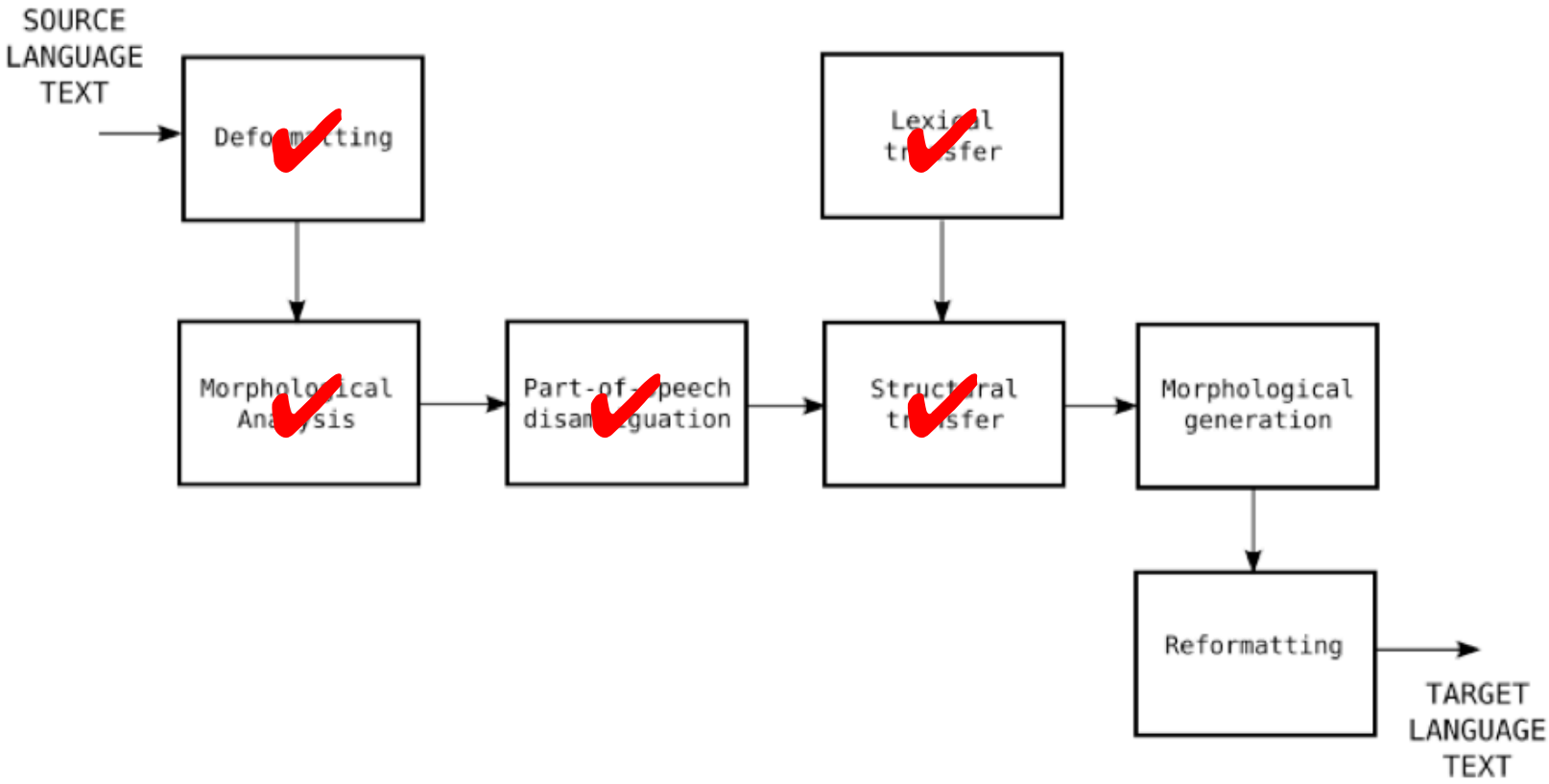
- Neste passo pode acontecer: inserção, eliminação, substituição, reordenação ou operações combinadas.

- **Exemplo:**

  - Palavra que numa língua é feminina e noutra masculina, sendo necessário mudar a própria palavra e/ou outras associadas;

    - ‘Vou ligar’ -> ‘Voy a conectar’

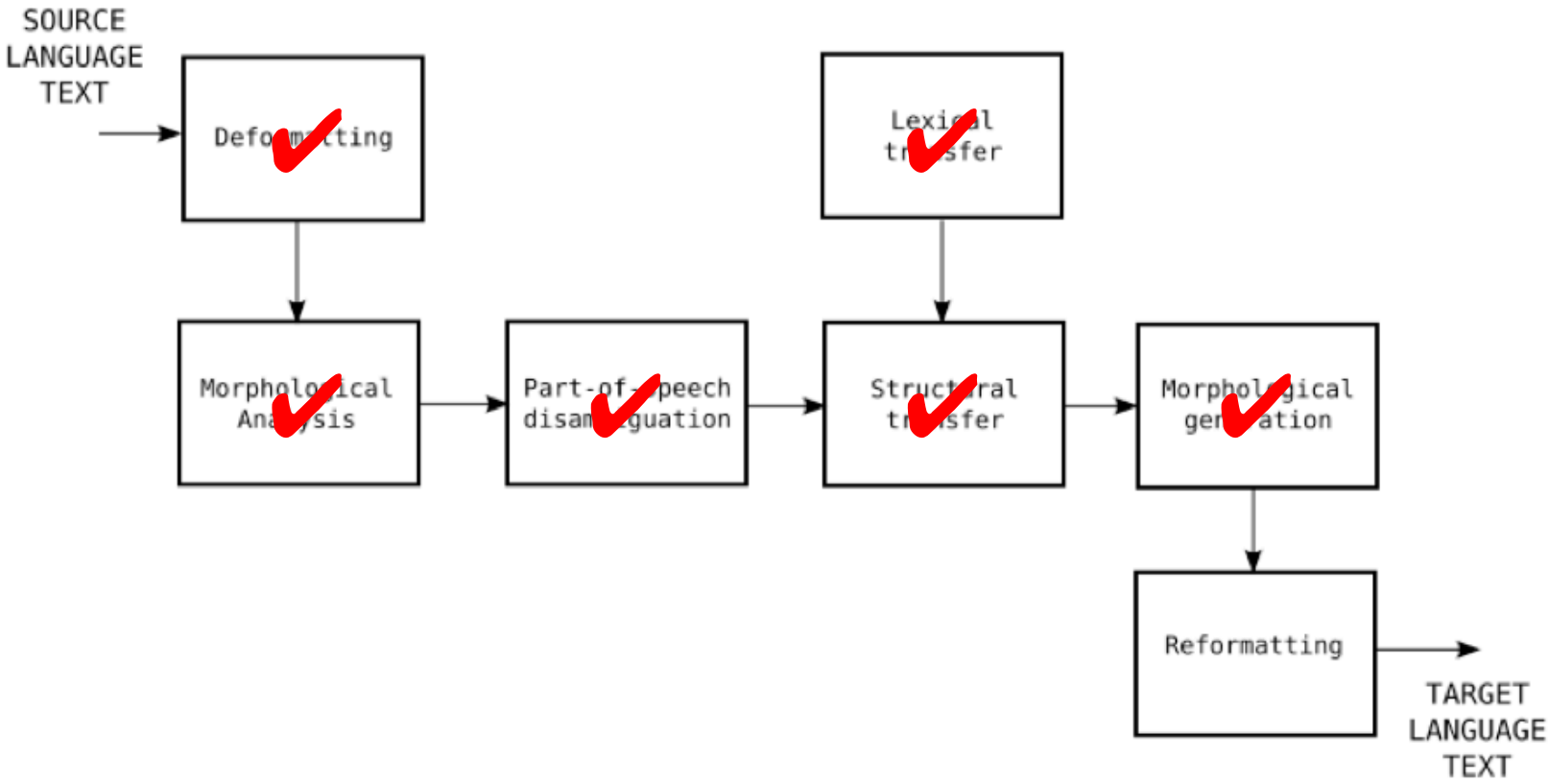
Como funciona?



Morphological Generation

- Neste passo passa-se o inverso da **Análise Morfológica**. É usado o dicionário da língua alvo e a partir do *lema* e das *tags*, as palavras são reconstruídas na língua final.

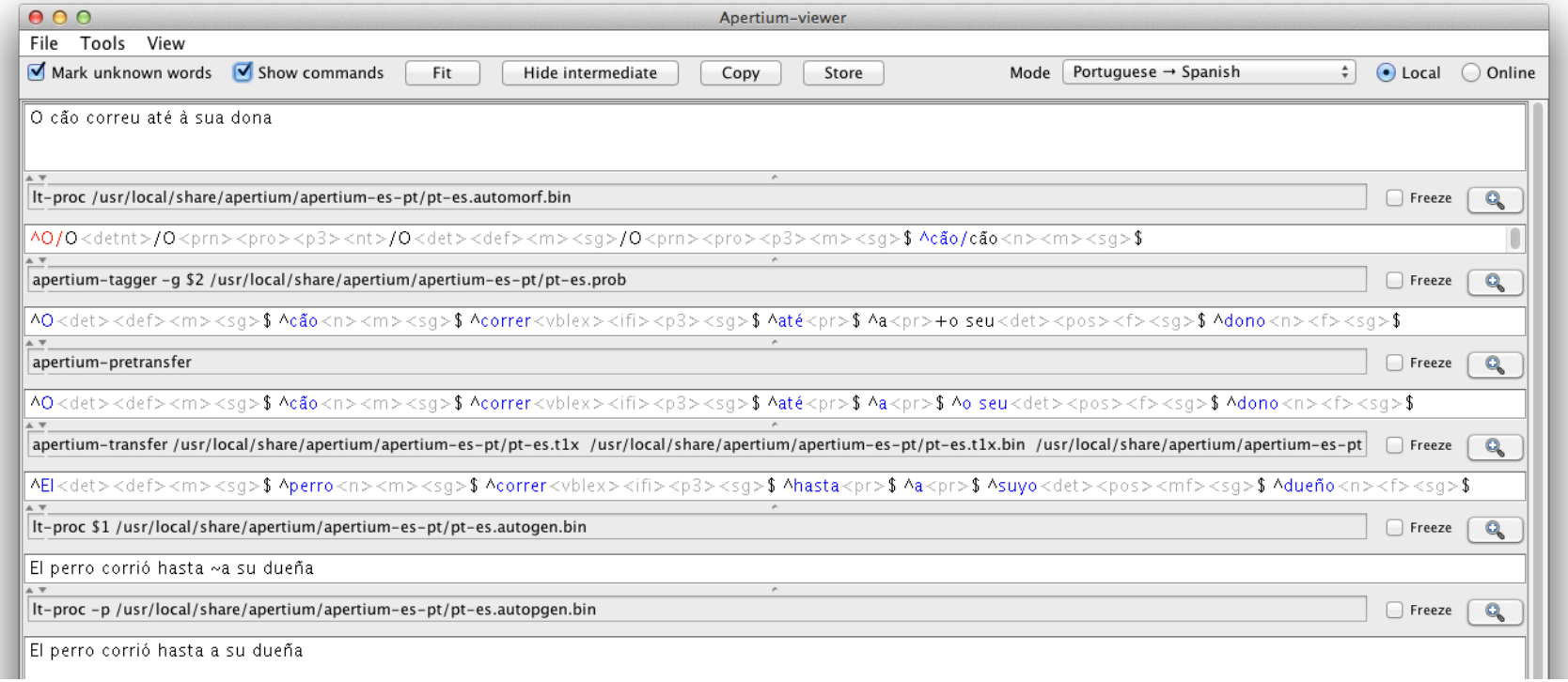
Como funciona?



Reformatting

- Inverso do primeiro passo. Restaura a informação referente ao formato do ficheiro, retirando os caracteres que a delimitavam;

Apertium - viewer



< prn > - pronome
< tn > - tónico
< nt > - neutro
< adv > - adverbio

< vbser > - verbo ser
< pri > - presente do indicativo
< p3 > - terceira pessoa
< f > - feminino

< sg > - singular
< det > - determinador
< ind > - indefinido
< sg > - singular
< n > - nome

```
$ echo 'Alguna cuestión?' | apertium es-pt
Alguma questão?
```