Conteúdo

1 Diretrizes para o alinhamento lexical **1**

1.1 Introdução . . . 1

1.2 Notação . . . 3

1.3 Regras gerais . . . 4

1.3.1 Alinhamento entre palavras com gênero e número distintos . . . 7

1.3.2 Alinhamento entre determinantes . . . 7

1.3.3 Preposição e artigo concatenados em uma mesma palavra . . . 7

1.3.4 Pronomes relativos . . . 8

1.3.5 Sintagma nominal X Sintagma verbal . . . 9

1.3.6 *Phrasal verbs* e verbos preposicionados . . . 9

1.3.7 Verbos principal e auxiliar(es) . . . 10

1.3.8 Verbo + "*se*" . . . 11

1.3.9 Substantivo composto . . . 11

1.3.10 Expressões fixas . . . 12

1.3.11 Outros alinhamentos envolvendo multipalavras . . . 14

1.3.12 Anáfora . . . 15

1.3.13 Seqüência de tokens repetida em apenas um dos lados . . . 15

1.4 Regras específicas para `PT-ES` . . . 16

1.4.1 Verbo + *la,lo,las,los,le,*etc. . . . 16

1.5 Regras específicas para `PT-EN` . . . 16

1.5.1 Preposição entre substantivos . . . 16

1.5.2 Possessivo . . . 17

1.5.3 Verbo com e sem sujeito . . . 17

1.6 Discondâncias . . . 18

1.7 Conclusões . . . 19

Resumo

Este relatório apresenta as diretrizes definidas, no âmbito do projeto ReTraTos, para o processo de alinhamento lexical de textos paralelos escritos em Português do Brasil, Espanhol e Inglês. Os textos paralelos e sua versão alinhada são, ambos, de grande importância para diversas aplicações de Processamento de Línguas Naturais (PLN), como: aprendizado de regras de tradução (objetivo do projeto ReTraTos), *Example-Based Machine Translation* (EBMT), *Statistical Machine Translation* (SMT), extração de léxicos bilíngües, desambiguação lexical de sentido, entre outras. As diretrizes aqui apresentadas possibilitaram a criação de córpus paralelos alinhados lexicalmente que seguem padrões bem definidos eliminando, assim, um grande número de ambigüidades inerentes do processo de alinhamento. Tanto os córpus quanto as diretrizes produzidos neste trabalho poderão ser utilizados em projetos futuros para a produção de ferramentas e recursos para o Processamento de Linguagem Natural.

Capítulo 1

Diretrizes para o alinhamento lexical

1.1 Introdução

Este relatório apresenta as diretrizes seguidas no processo de marcação manual de alinhamentos lexicais em textos paralelos escritos em Português do Brasil (PT), Inglês (EN) e Espanhol (ES) realizada no âmbito do projeto ReTraTos.[1] Os textos paralelos alinhados lexicalmente são de grande importância para diversas aplicações, como: aprendizado de regras de tradução [Carl, 2001, Menezes and Richardson, 2001] – objetivo do projeto ReTraTos – , *Example-Based Machine Translation* (EBMT) [Somers, 1999], *Statistical Machine Translation* (SMT) [Ayan et al., 2004, Och and Ney, 2000], extração de léxicos bilíngües [Melamed, 1996, Gómez Guinovart and Sacau Fontenla, 2004], desambiguação lexical de sentido [Gale et al., 1992], entre outras.

Nos últimos anos, diversos métodos automáticos de alinhamento lexical têm sido propostos na literatura [Hiemstra, 1998, Och and Ney, 2000, Ayan et al., 2004, Wu and Wang, 2004] com resultados que variam de 71% a 92% de precisão e 62% a 88% de cobertura de acordo com o método e o par de línguas envolvidas, entre outros fatores. Porém, para calcular automaticamente essas métricas é necessário dispor de um corpus de referência, ou seja, um conjunto de textos paralelos alinhados lexicalmente por um especialista humano e, portanto, considerados corretos. Desta maneira, a precisão e a cobertura são calculadas comparando-se os alinhamentos gerados pelo método automático com aqueles do corpus de referência.[2]

Neste sentido, este relatório apresenta as principais diretrizes definidas no processo de marcação manual de alinhamentos entre palavras e multipalavras (conjunto de duas ou mais

[1]O projeto ReTraTos é desenvolvido com o apoio de FAPESP, CAPES e CNPq.

[2]A precisão é calculada como o número de tokens (palavras, números, etc.) fonte e alvo em comum nos alinhamentos proposto (pelo método automático) e de referência, dividido pelo número de tokens no alinhamento proposto; enquanto cobertura é o número de tokens fonte e alvo em comum nos alinhamentos proposto e de referência, dividido pelo número de tokens no alinhamento de referência.

palavras) para a criação de córpus de referência com textos paralelos PT-ES e PT-EN. Exemplos de projetos que também definiram suas regras para o processo de marcação de alinhamentos lexicais são o projeto ARCADE [3] e o projeto Blinker [Melamed, 1998], nas quais este relatório está fortemente baseado. Os córpus de referência gerados como produto do alinhamento lexical dos textos paralelos PT-ES e PT-EN foram usados na avaliação dos alinhamentos gerados automaticamente pelo alinhador lexical LIHLA criado no âmbito do projeto ReTraTos.[4]

As regras definidas para o processo de marcação manual de alinhamentos lexicais em 14 pares de textos paralelos PT-ES foram obtidas em um processo de 3 passos:

1. Primeiro, dois anotadores nativos do PT e com conhecimentos em ES anotaram, separadamente, todos os tokens (palavras, caracteres de pontuação e números) de todas as sentenças dos 14 pares de textos paralelos PT-ES;

2. Em seguida, as anotações foram comparadas e os casos discrepantes foram discutidos e ajustados. Uma taxa de concordância de 95% entre os dois anotadores – número de alinhamentos em comum dividido pelo número total de alinhamentos produzidos – foi obtida considerando-se, apenas, alinhamentos extatamente iguais, ou seja, concordâncias parciais foram consideradas como alinhamentos discordantes.

3. Uma versão final do corpus lexicalmente alinhado, bem como, das regras utilizadas para a criação desta anotação foi compilada gerando como co-produto este relatório.

A partir das regras para PT-ES, 10 pares de textos paralelos PT-EN foram alinhados lexicalmente a fim de verificar se as regras também eram válidas para este par de línguas. Além disso foram adicionadas novas regras para casos específicos não contemplados anteriormente. Assim, no total foram alinhados 15.396 tokens no corpus PT-ES (7.236 tokens em PT e 8.160 tokens em ES) e 15.900 no corpus PT-EN (7.631 tokens em PT e 8.269 tokens em EN).

Assim, este relatório está organizado como se segue. A primeira seção (1.2) explica a notação utilizada neste documento para descrever as diretrizes do processo de alinhamento lexical manual. Na seção 1.3, são apresentadas as regras gerais aplicáveis tanto aos textos PT-ES quanto aos textos PT-EN; enquanto que as seções 1.4 e 1.5 apresentam as regras específicas para os pares de línguas PT-ES e PT-EN, respectivamente. Alguns exemplos de alinhamentos nos quais

[3]Cujos critérios para marcação manual de alinhamentos lexicais está disponível em: http://www.up.univ-mrs.fr/veronis/arcade/arcade1/2nd/word/guide/index.html.

[4]Para mais informações sobre projetos e recursos do NILC, consulte o *web-site* do laboratório: http://www.nilc.icmc.usp.br.

os anotadores discordaram são mostrados na seção 1.6 e, por fim, na seção 1.7 são apresentadas algumas conciderações finais a respeito deste trabalho.

1.2 Notação

Antes de apresentar as regras especificadas para o processo de alinhamento lexical, bem como os exemplos reais encontrados nos textos paralelos `PT-ES` e `PT-EN`, é necessário apresentar algumas considerações a respeito da notação utilizada neste documento.

- **Tipo do alinhamento lexical**

  O tipo de um alinhamento lexical é indicado por uma seqüência $l : d$, onde l e d são número inteiros maiores ou iguais a 0 que indicam a quantidade de tokens fonte (do lado esquerdo) e alvo (do lado direito). Embora o tipo mais comum de alinhamento seja o $1 : 1$ (no qual um token no texto fonte é traduzido exatamente como um token no texto alvo), outros tipos de alinhamentos como omissões ($1 : 0$ ou $0 : 1$) ou os que envolvem multipalavras também são possíveis. Exemplos de tipos de alinhamentos envolvendo multipalavras são: expansões ($n : m$, com $n < m; n, m \geq 1$), contrações ($n : m$, com $n > m; n, m \geq 1$) e uniões ($n : n$, com $n > 1$).

- **Alinhamento**

  Um alinhamento é indicado, no escopo deste documento, como uma seqüência de um ou mais tokens fonte (ou simplesmente a palavra NULL), um caractere "⇔" e uma seqüência de tokens alvo (ou NULL). Quando a posição em que um token aparece no texto for importante para o entendimento do contexto no qual uma regra deve ser aplicada, sua posição será indicada por um número seguido do caractere ":" precedendo o token. Por exemplo, em $1{:}X \Leftrightarrow 2{:}\,Y$, o token X que aparece na posição 1 do texto fonte está alinhado com o token que aparece na posição 2 do texto alvo (Y).

- **NULL**

  A palavra especial NULL indica um alinhamento de omissão ($1 : 0$ ou $0 : 1$) e deve ser utilizada para alinhar as palavras que não possuem correspondência no texto paralelo.

- **Categorias gramaticais**

  Com o intuito de tornar a aplicação de uma regra o mais geral possível, às vezes, na sua definição indica-se a categoria gramatical a qual um token deve pertencer. Assim, as

categorias gramaticais utilizadas neste documento são as seguintes: PREP (preposição), VERB (verbo), AUX (verbo auxiliar), SUBS (substantivo), ART (artigo), PRON (pronome).

- **Caracteres com significados especiais**

  [ ] - Os colchetes indicam que a presença do token delimitado por eles é opcional. Assim, por exemplo a seqüência [PREP]VERB indica a presença de um verbo que pode ou não vir precedido por uma preposição.

  _ - O caractere "_" entre dois tokens indica que eles estão concatenados e, portanto, são representados em apenas um token. Por exemplo, a seqüência PREP_ART indica que uma preposição e um artigo são encontrados conjuntamente em um mesmo token, como é o caso de "*dos*"em PT ou "*al*" em ES.

  | - O caractere "|" indica ou lógico, ou seja, um (e apenas um) dos tokens separados por "|" poderá ocorrer. Por exemplo, [PREP|ART] indica que se ocorrer (os colchetes indicam que são opcionais) ocorrerá apenas uma preposição ou um artigo, nunca os dois ao mesmo tempo.

  + - O caractere "+" indica a união de tokens para formar um alinhamento $n : m$ com n e/ou $m > 1$, ou seja, um alinhamento envolvendo uma ou mais multipalavras. Por exemplo, seria necessário unir a preposição "*of*" e o artigo "*the*" do EN para alinhá-los com PREP_ART "*da*" do PT, gerando un alinhamento $2 : 1$ (no sentido EN-PT).

1.3 Regras gerais

Assim, o processo de marcação manual de alinhamentos entre palavras e multipalavras pode ser efetuado seguindo 3 passos:

Para cada token (palavra, número, caractere de pontuação):

1. Busque o melhor token do outro lado cujo significado é o mesmo do token sendo alinhado. Se a correspondência um-para-um se der nos dois sentidos (fonte-alvo e alvo-fonte) gere um alinhamento $1 : 1$, senão vá para o passo 2.
2. Busque o menor conjunto de tokens no outro lado com o mesmo significado do token sendo alinhado, produzindo um alinhamento

$1:n$ (ou $n:1$). Porém, se forem necessários mais de um token em ambos os lados para expressar o mesmo significado, então vá para o passo 3.

3. Tente gerar subalinhamentos envolvendo os menores conjuntos de tokens em ambos os lados, sempre garantindo a equivalência semântica. Porém, se não for possível gerar vários subalinhamentos ou houver dúvidas a respeito da equivalência semântica, gere apenas um alinhamento $n:m$ envolvendo todos os n tokens em um lado e os m tokens no outro.

Assim, algumas regras gerais podem ser derivadas da seqüência de passos apresentada anteiormente.

R1. Crie o menor alinhamento que mantenha o significado nos dois sentidos (fonte-alvo e alvo-fonte).

Exemplo (PT-ES):

1. "*sem ultrapassar o*" e "*que no irán más allá del*"
   Neste caso pode-se identificar dois blocos de alinhamentos *sem+ultrapassar⇔que+no+(va)+más+allá+de* e *o⇔el*, mas como não é possível separar a preposição do artigo na palavra "*del*", os dois blocos devem ser unidos em um único alinhamento:
   sem+ultrapassar+o⇔que+no+irán+más+allá+del

2. "*operadores de direito*" e "*legal workers*"
   Neste caso pode-se identificar dois blocos de alinhamentos:
   operadores⇔workers
   de+direito⇔legal

R2. Especifique a correspondência mais detalhada possível; porém, se houver dúvidas a respeito de como gerar subalinhamentos, gere um único alinhamento englobando todos os tokens.

Exemplo (PT-EN):

1. "*estimula*" e "*is capable of stimulating*"
   Neste caso não é possível gerar um alinhamento distinto de:
   estimula⇔is+capable+of+stimulating

2. "*ao que tudo indica*" e "*or so everything indicates*"

   Neste caso não é tão fácil determinar os alinhamentos na primeira análise, porém pode-se identificar dois alinhamentos 1 : 1: um entre "*indica*" e "*indicates*" e outro entre "*tudo*" e "*everything*". Há também uma possibilidade, mais fraca, de estabelecer uma correspondência entre "*ao que*" e "*or so*". Assim, opta-se pelos alinhamentos:

   ao+que⇔*or+so*

   tudo⇔*everything*

   indica⇔*indicates*

R3. Um alinhamento de um token com NULL (um caso de omissão) deve ser gerado quando o token que parece estar "sobrando" não puder ser adicionado ao token que vem antes ou depois porque não é essencial para seu significado e, em alguns casos, isto até prejudicaria o alinhamento já existente entre os tokens vizinhos.

Exemplo (`PT-EN`):

1. "*o pau-brasil*" e "*brasilwood*"

   Neste caso não se deve adicionar o artigo definido do `PT` "*o*" ao alinhamento de "*pau-brasil*" com "*brazilwood*", pois isso prejudicaria a correspondência existente entre esses tokens. Desta maneira, o correto seria gerar dois alinhamentos:

   o⇔NULL

   pau-brasil⇔*brazilwood*

2. "*rapidamente*"e "*con maior rapidez*"

   Neste exemplo percebe-se que as duas palavras em `ES` "*con rapidez*" possuem o mesmo significado da palavra em `PT` "*rapidamente*", porém a palavra "*maior*" não possui correspondência no outro lado, assim dois alinhamentos devem ser gerados:

   rapidamente⇔*con+rapidez*

   NULL⇔*maior*

R4. Os caracteres de pontuação devem ser alinhados de uma maneira que garanta um número mínimo de links cruzados e é permitido alinhar caracteres diferentes entre si e até mesmo caracteres de pontuação com palavras, por exemplo "*e*⇔," e "*y*⇔;".

Exemplo (`PT-ES`):

1. "*a forma mais branda, a cutânea,*" e "*la forma cutánea, la más benigna,*"

   Neste caso, os alinhamentos entre os caracteres de pontuação "," que geram o menor número de cruzamentos é obtido ao alinhá-los na ordem em que aparecem.

A seguir são apresentados alguns casos que merecem maior atenção por parte do anotador.

1.3.1 Alinhamento entre palavras com gênero e número distintos

As palavras com gênero e/ou número distintos podem ser alinhadas entre si contanto que o significado seja preservado.

Exemplos (PT-EN):

- *metro*⇔*meters*
- *comuns*⇔*common*
- *florestas*⇔*forest*

Exemplos (PT-ES):

- *no*⇔*en+la*
- *estoque*⇔*existencias*
- *total*⇔*totales*

1.3.2 Alinhamento entre determinantes

Os determinantes podem ser alinhados entre si mesmo que não pertençam a mesma categoria gramatical.

Exemplos (PT-EN):

- *os*⇔*those* (ART⇔PRON)
- *uma*⇔*its* (ART⇔PRON)

Exemplos (PT-ES):

- *o*⇔*ese* (ART⇔PRON)
- *o*⇔*Este* (ART⇔PRON)

1.3.3 Preposição e artigo concatenados em uma mesma palavra

Quando, de um lado, houver uma concatenação de PREP_PRON ou PREP_ART em um só token e, do outro, houver apenas uma das partes (PREP ou PRON ou ART) o alinhamento é feito 1 : 1, mesmo que se refira a apenas uma parte, pois não é possível dividir a concatenação.

Exemplos (PT-EN):

- *da*⇔*of* (PREP_ART⇔PREP)
- *desse*⇔*this* (PREP_PRON⇔PRON)
- *dos*⇔*the* (PREP_ART⇔ART)

Exemplos (PT-ES):

- *dos*⇔*de* (PREP_ART⇔PREP)
- *o*⇔*al* (ART⇔PREP_ART)
- *no*⇔*en* (PREP_ART⇔PREP)

Porém, NÃO se deve gerar alinhamentos do tipo $1:n$ ou $n:1$ envolvendo PREP, ART e PRON quando houver a possibilidade de alinhar $1:1$ (PREP⇔PREP ou ART⇔ART ou PRON⇔PRON). Neste caso, as outras partes se mantém não-alinhadas.

Exemplos (PT-EN):

- Não faça: *de*⇔*of+the*

  Faça: *de*⇔*of* e NULL⇔*the*

Exemplos (PT-ES):

- Não faça: *de*⇔*de+la*

  Faça: *de*⇔*de* e NULL⇔*la*

Além disso, quando, de um lado, houver uma concatenação PREP_PRON ou PREP_ART em um só token e, do outro, houver as duas partes separadas (PREP e PRON ou PREP e ART), essas devem ser unidas para garantir 100% da tradução.

Exemplos (PT-EN):

- *pela*⇔*by+the* (PREP_ART⇔PREP+ART)
- *dos*⇔*of+those* (PREP_ART⇔PREP+PRON) (veja regra 1.2.2)
- *desse*⇔*of+this* (PREP_PRON⇔PREP+PRON)

Exemplos (PT-ES):

- *ao*⇔*a+lo* (é possível) (PREP_ART⇔PREP+ART)
- *desse*⇔*de+ese* (PREP_PRON⇔PREP+PRON)
- *na*⇔*en+su* (PREP_ART⇔PREP+PRON) (veja regra 1.2.2)

1.3.4 Pronomes relativos

Quando o alinhamento envolver pronomes relativos, deve-se optar sempre pelo alinhamento mais abrangente.

Exemplos (PT-EN):

- *as+quais*⇔*which*
- *que*⇔*which*
- *em+que*⇔*where*
- *de+que*⇔*than*

Exemplos (PT-ES):

- *as⇔las*

  quais⇔cuales

- *a⇔a*

  que⇔la+cual

- *en⇔em*

  que⇔los+que

- *do+que⇔que*

1.3.5 Sintagma nominal X Sintagma verbal

Quando, de um lado, houver um sintagma nominal e, do outro, um sintagma verbal, todos os tokens que formam os sintagmas (PREP, ART, SUBS, AUX, VERB), devem ser unidos para gerar um alinhamento 1 : 1 entre os sintagmas.

Exemplos (PT-EN):

- *ao+vôo⇔to+fly*
- *à+eliminação⇔to+putting+down*

Exemplos (PT-ES):

- *ao+vôo⇔a+volar*
- *tratar⇔el+tratamiento*

1.3.6 *Phrasal verbs* e verbos preposicionados

A preposição deve ser concatenada ao verbo em um ou ambos os lados sempre que ela fizer parte do significado de tal verbo. Neste sentido, os *phrasal verbs* do inglês nunca devem ser separados, mesmo que seja necessário adicionar uma preposição ao verbo no outro idioma.

Exemplos (PT-EN):

- *cuida+do⇔is+looking+to*

  Pois: *cuida+de⇔look+to*

- *realizadas⇔carried+out*
- *ativa⇔sets+off*
- *fazem+parte⇔make+up+part*
- *deparam+com⇔come+across*

Exemplos (PT-ES):

- *fazem+parte⇔formam+parte*

- *é*⇔*consiste+en*

1.3.7 Verbos principal e auxiliar(es)

Os verbos auxiliares não devem ser unidos ao verbo principal quando possuírem correspondência do outro lado.

Exemplos (PT-EN):

- *não*⇔*not*

  tenha⇔*has*

  chegado⇔*reached*
- *haviam*⇔*had*

  sido⇔*been*

  contaminados⇔*contamined*
- *vem*⇔*has+been*

  aumentando⇔*increasing*

Exemplos (PT-ES):

- *não*⇔*no*

  tenha⇔*ha*

  chegado⇔*llegado*
- *vem*⇔*ha+ido*

  aumentando⇔*en+aumento*

Porém, quando houver auxiliar(es) em um lado mas não do outro, ambos os verbos auxiliar(es) e principal devem ser alinhados com o verbo principal do outro lado. O mesmo se aplica a outras partículas (diferentes de verbos auxiliares) que são essenciais para garantir a equivalência semântica, principalmente com verbos na voz passiva (veja os últimos exemplos a seguir).

Exemplos (PT-EN):

- *acompanharão*⇔*will+accompany*
- *serão*⇔*will+be*
- *identificou*⇔*has+identified*
- *testa*⇔*is+testing*
- *não*⇔*not*

  surte⇔*does+produce*
- *subiram*⇔*were+launched*

- *se+reduzir⇔to+be+reduced*

Exemplos (PT-ES):

- *vão+integrar⇔integrarán*
- *iniciam⇔dan+inicio*
- *testa⇔está+probando*
- *registraram⇔han+registrado*
- *subiram⇔se+lanzaron*
- *vem+sendo+aplicado⇔se+lo+está+aplicando*

1.3.8 Verbo + "*se*"

Quando o token "*se*" em PT ou ES faz parte do significado de um verbo, deve ser unido ao verbo (caso contrário deve permanecer não-alinhado). Além disso, quando o token "*se*" estiver concatenado ao verbo em espanhol, ele deve ser unido ao verbo em PT para garantir o mesmo significado na tradução (exemplos 3 e 4 a seguir).

Exemplos (PT-EN):

- *torna+-+se⇔becomes*
- *vão⇔are+going+to*

  se+integrar⇔join

Exemplos (PT-ES):

- *cuida⇔se+encarga*
- *deparam⇔se+encuentram*
- *se+reduzir⇔reducirse*
- *se+tornar⇔convertirse*

1.3.9 Substantivo composto

Quando houver apenas um SUBS de um lado e do outro lado vários tokens essenciais para garantir o significado deste substantivo, então todos os tokens deste outro lado devem ser unidos em um único alinhamento, mesmo que uma correspondência 1 : 1 possa ser estabelecida entre apenas um desses tokens e o SUBS do outro lado. Isto porque, os tokens que parecem "sobrar" do outro lado são essenciais para o entendimento da idéia, naquele idioma.

Exemplos (PT-EN):

- *redes+de+arrasto⇔dragnets*
- *batimentos+cardíacos⇔heartbeats*

- *namorado⇔namorado+sandperch*
- *batata⇔potato+-+fish*
- *Bauru⇔the+town+of+Bauru*
- *Minas⇔the+state+of+Minas+Gerais*
- *endemias⇔endemic+diseases*
- *médias⇔medium+sized*

  e⇔and

  grandes⇔large+sized

  cidades⇔cities

Exemplos (`PT-ES`):

- *cherne⇔cherna+pinta*
- *batata⇔pez+ batata*
- *Minas⇔Minas+Gerais*
- *Maranhão⇔estado+de+Maranhão*
- *site⇔sitio+en+internet*
- *sudeste⇔región+sudeste*

1.3.10 Expressões fixas

Os tokens que formam uma expressão fixa de um lado devem ser sempre unidos mesmo que para garantir o significado seja necessário unir vários tokens em ambos os lados. Além disso, as expressões fixas devem ser alinhadas mesmo que não ocorram em posições consecutivas no texto (como o último exemplo de `PT-EN` apresentado a seguir).

Exemplos (`PT-EN`):

- *ao+redor+da⇔around+the*

  Pois: *ao+redor+de⇔around*
- *de+acordo+com⇔according+to*
- *ao+lado+de⇔next+to*
- *tal+qual⇔just+like*
- *a+partir+dessa⇔following+this*

  Pois: *a+partir+de⇔following*
- *segundo⇔according+to*
- *em+conjunto⇔as+a+whole*
- *na+casa+dos⇔in+the+region+of*

- *por+meio+de⇔by+means+of*
- *por+ora⇔for+the+time+being*
- *em+curso⇔under+way*
- *Comunidade+Européia⇔European+Union*
- 65:*tanto*+69:*quanto*⇔94:*both*+97:*and*

  66:*a*⇔NULL

  67:*leishmaniose*⇔96:*leishmaniasis*

  68:*tegumentar*⇔95:*tegumentary*

  70:*a*⇔98:*the*

  71:*visceral*⇔99:*visceral*

Exemplos (PT-ES):

- *de+acordo+com⇔conforme*
- *de+acordo+com⇔según*
- *ao+lado+de⇔junto+a*
- *tal+qual⇔a+la+manera+de*
- *a+partir+dessa⇔con+base+en+esta*

  Pois: *a+partir+de⇔con+base+en*
- *além+de⇔al+margem+de*
- *abaixo⇔[por+]debajo*
- *em+conjunto⇔conjuntamente*
- *na+casa+dos⇔alrededor+de*
- *embora⇔pese+a[+que]*
- *em+vez+de⇔en+lugar+de*
- *por+meio+de⇔a+través+de*
- *mas⇔sino+también*

Deve-se, também, priorizar alinhamento mais abrangente, ou seja, se existir a possibilidade (definida em um dicionário bilíngüe como possível tradução) de unir mais de um token e manter o significado, eles devem ser unidos mesmo que apenas uma parte da locução já garanta o significado da tradução.

Exemplos (PT-EN):

- *desde[+que]⇔ever+since*
- *mais[+de]⇔over*
- *novamente⇔[once+]again*

Exemplos (PT-ES):

- *apenas⇔[tan+]solo*
- *sozinho⇔[por+sí+]solo*
- *de⇔[a+partir+]de*
- *de⇔[por+parte+]de*
- *via⇔[por+]vía*
- *até[+mesmo]⇔incluso*
- *hoje⇔hoy[+en+día]*
- *como⇔[tal+]como*

OBS.: Nome próprio

Porém, quando se tratar de nomes próprios e for possível estabelecer vários alinhamentos 1 : 1 entre seus tokens constituintes, deve-se respeitar o princípio do menor alinhamento que preserva a equivalência semântica (R1). Por exemplo:

- Não faça: *Belo+Horizonte⇔Belo+Horizonte*
  Faça: *Belo⇔Belo* e *Horizonte⇔Horizonte*
- Não faça: *Rio+de+Janeiro⇔Rio+de+Janeiro*
  Faça: *Rio⇔Rio* e *de⇔de* e *Janeiro⇔Janeiro*

1.3.11 Outros alinhamentos envolvendo multipalavras

Além dos exemplos de alinhamentos envolvendo multipalavras apresentados anteriormente, muitos outros são possíveis e devem ser criados tendo como base as regras gerais apresentadas no início desta seção.

Exemplos (PT-EN):

- *dominó⇔dominoes*
  que+tomba⇔falling
- *mosquito⇔mosquito*
  transmissor⇔that+transmits
  causador⇔that+causes
- *apoio⇔support*
  dos+técnicos⇔technical
- *equipe⇔team*
  baiana⇔from+Bahia

Exemplos (PT-ES):

- *esquecidas*⇔*que+estaban+olvidadas*
- *prejudicando*⇔*que+prejudica*
- *emergenciais*⇔*de+emergencia*
- *consorciada*⇔*por+vía+de+consorcios*
- *de+moradores+de+bairro*⇔*vecinales*
- *decisórios*⇔*de+decisión*

1.3.12 Anáfora

Permite-se alinhar dois tokens que não sejam tradução um do outro se o papel que desempenham no texto é o mesmo e este papel não puder ser desempenhado por nenhum outro token no contexto.

Exemplos (`PT-EN`):

- *as+plantas*⇔*they*
- *os+pesquisadores*⇔*they*

Exemplos (`PT-ES`):

- *nesses*⇔*de+dichos*
- *peixe*⇔*especie*

1.3.13 Seqüência de tokens repetida em apenas um dos lados

Às vezes, pode ocorrer que em um dos lados uma seqüência de tokens apareça mais de uma vez, mas que na tradução exista apenas uma ocorrência. Neste caso, a seqüência que ocorre apenas uma vez deve ser alinhada com todas as ocorrências do outro lado. Assim, um ID que aparece mais de uma vez em alinhamentos diferentes não indica alinhamento de multipalavras, mas sim vários alinhamentos possíveis para aquele ID.

Exemplos (`PT-EN`):

- 5:*do*⇔**5**:*of*+**6**:*the*
  nariz⇔*nose*
  ,⇔,
  8:*da*⇔**5**:*of*+**6**:*the*
  boca⇔*mouth*
  e⇔*and*
  11:*da*⇔**5**:*of*+**6**:*the*
  garganta⇔*throat*

Exemplos (PT-ES):

- 5:*do*⇔**5**:*de*+6:*la*

  nariz⇔*nariz*

  ,⇔,

  8:*da*⇔**5**:*de*+9:*la*

  boca⇔*boca*

  e⇔*y*

  11:*da*⇔**5**:*de*+12:*la*

  garganta⇔*garganta*

1.4 Regras específicas para PT-ES

1.4.1 Verbo + *la,lo,las,los,le,*etc.

Uma das características da língua espanhola é a concatenação de partículas ao verbo quando este é apresentado no infinitivo, gerúndio ou imperativo. Assim, quando um verbo no espanhol contiver uma destas partículas concatenadas, deve-se unir ao verbo do outro lado quantos tokens forem necessários para garantir o significado.

Exemplos (PT-ES):

- *regular+essa+resposta*⇔*regularla*
- *tornam*⇔*continúan+tornándolo*
- *armazená+-+lo*⇔*almacenarlas*

1.5 Regras específicas para PT-EN

1.5.1 Preposição entre substantivos

Este é um caso bastante frequente na tradução PT-EN e deve ser tratado seguindo a regra **R3** (não se deve adicionar a PREP ao alinhamento dos substantivos que a rodeiam).

Exemplos (PT-EN):

- *vapor*⇔*vapor*

  de⇔NULL

  água⇔*water*

- *Instituto*⇔*Institute*

  de⇔NULL

  Pesca⇔*Fishing*

- *testes*⇔*tests*

  em⇔NULL

  campo⇔*field*

- *semente*⇔*seed*

  do⇔NULL

  pau+-+brasil⇔*brazilwood*

1.5.2 Possessivo

A partícula indicativa de posse do inglês – *'s* ou apenas *'* – deve ser alinhada à preposição (possivelmente concatenada a um artigo) que desempenha o mesmo papel no português, quando esta estiver presente.

Exemplos (`PT-EN`):

- *do*⇔*'s*

  organismo⇔*organism*

- *dos*⇔*'*

  organismos⇔*organisms*

- *Alzheimer*⇔*Alzheimer*

  NULL⇔*'s*

1.5.3 Verbo com e sem sujeito

Se o sujeito do verbo em `PT` aparecer apenas implicitamente na terminação verbal então o sujeito do verbo em `EN` (se estiver explicitamente definido) deve, necessariamente, ser unido ao verbo para estabelecer o alinhamento mesmo que o sujeito e o verbo não apareçam em posições consecutivas no texto.

Exemplos (`PT-EN`):

- *atravessaram*⇔*they+have+crossed+out*
- *Vimos*⇔*We+saw*
- *duravam*⇔*they+would+last*
- *fossem*⇔*they+were*
- 5:*às*+6:*vezes*⇔6:*sometimes*

  7:*tenho*⇔5:*I*+7:*have*

Porém, se o sujeito não aparece explicitamente e também não pode ser identificado na terminação do verbo, então deve-se seguir a regra do menor alinhamento e alinhar apenas os

verbos entre si.

Exemplos (PT-EN):

- NULL⇔*it*

  destruir⇔*destroys*

- NULL⇔*when*

  NULL⇔*they*

  seguidos⇔*followed*

- NULL⇔*it*

  podendo⇔*can*

1.6 Discondâncias

Nesta seção são apresentados alguns casos em que as anotadoras discordaram na criação dos alinhamentos.

Tabela 1.1: Exemplos de discordâncias entre os anotadores

Anotador A	**Anotador B**
a+exemplo+do⇔*como+por+ejemplo+el*	*a*⇔*por* *exemplo*⇔*ejemplo* *do*⇔*como+el*
menos⇔*de+menor* *nobre*⇔*calidad*	*menos+nobre*⇔*de+menor+calidad*
tornar⇔*convierta* NULL⇔*en*	*tornar*⇔*convierta+en*
de+modo⇔*en+forma*	*de*⇔*en* *modo*⇔*forma*
deu⇔*verificó* *apenas*⇔*en+tan+solo*	*deu*⇔*verificó+en* *apenas*⇔*tan+solo*
demonstrar⇔*hacer* *publicamente*⇔*[75]:público*	*demonstrar+publicamente*⇔*hacer+público*
como⇔*al+modo*	*como*⇔*al+modo+de*
lei⇔*ordenanza+municipal*	*lei*⇔*ordenanza* NULL⇔*municipal*
virou⇔*se+ha+convertido+en*	NULL⇔*se* *virou*⇔*ha+convertido+en*
NULL⇔*en* NULL⇔*medio* NULL⇔*a* *num*⇔*un*	*num*⇔*en+medio+a+un*

1.7 Conclusões

O alinhamento lexical de palavras e multipalavras é um processo trabalhoso uma vez que, na maioria dos casos, não é possível estabelecer um alinhamento um-para-um para todos os tokens de um par de textos paralelos devido tanto a divergências entre as línguas quanto a liberdade nas traduções.

Assim, neste relatório foram apresentadas as principais diretrizes definidas para o processo de alinhamento lexical manual desempenhado no âmbito do projeto ReTraTos para a criação de córpus de referência alinhados lexicalmente envolvendo os pares de línguas `PT-ES` e `PT-EN`.

A partir dessas diretrizes, foi possível criar córpus de referência que seguem um padrão bem definido para o alinhamento lexical, eliminando assim grande número de possíveis ambigüidades. Os córpus produzidos como resultado deste processo, juntamente com as diretrizes aqui apresentadas, poderão ser utilizados em projetos futuros para a produção de ferramentas e recursos para o Processamento de Língua Natural.

Bibliografia

[Ayan et al., 2004] Ayan, N. F., Dorr, B. J., and Habash, N. (2004). Multi-Align: Combining linguistic and statistical techniques to improve alignments for adaptable MT. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 17–26. Springer-Verlag Berlin Heidelberg.

[Carl, 2001] Carl, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of CoNLL-2001*, pages 145–151, Toulouse, France.

[Gale et al., 1992] Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1992)*, pages 101–112, Montreal, Canada.

[Gómez Guinovart and Sacau Fontenla, 2004] Gómez Guinovart, X. and Sacau Fontenla, E. (2004). Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural*, 33:133–140.

[Hiemstra, 1998] Hiemstra, D. (1998). Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. In Coppen, P. A., van Halteren, H., and Teunissen, L., editors, *Proceedings of the 8th CLIN meeting*, pages 41–58.

[Melamed, 1996] Melamed, I. D. (1996). Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, pages 125–134, Montreal, Canada.

[Melamed, 1998] Melamed, I. D. (1998). Annotation style guide for the Blinker project, version 1.0.4. Technical Report 98-06, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA.

[Menezes and Richardson, 2001] Menezes, A. and Richardson, S. D. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pages 39–46, Toulouse, France.

[Och and Ney, 2000] Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 440–447, Hong Kong, China.

[Somers, 1999] Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

[Wu and Wang, 2004] Wu, H. and Wang, H. (2004). Improving domain-specific word alignment with a general bilingual corpus. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004)*, number 3265 in Lecture Notes in Artificial Inteligence (LNAI), pages 262–271. Springer-Verlag Berlin Heidelberg.