UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

Análise Comparativa de Tradutores Automáticos Inglês-Português

Marcela Franco Fossey
Tatiana Pedrolongo
Ronaldo Teixeira Martins
Maria das Graças Volpe Nunes

Nº 233

RELATÓRIOS TÉCNICOS

São Carlos – SP
Mai./2004

SYSNO 1388194
DATA / /
ICMC - SBAB

Universidade de São Paulo - USP

Universidade Federal de São Carlos - UFSCar

Universidade Estadual Paulista - UNESP

Análise comparativa de tradutores automáticos inglês-português

Marcela Franco Fossey

Tatiana Pedrolongo

Ronaldo Teixeira Martins

Maria das Graças Volpe Nunes

NILC-RT-04-04

Maio, 2004

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório visa descrever as atividades realizadas para testar e avaliar quatro ferramentas de Tradução Automática disponíveis na Web: Linguatec e-translation Server[1] (ES), Intertran[2] (IT), Systran [3](SY) e FreeTranslation [4](FT). Os testes foram feitos tendo como língua fonte o inglês e como língua meta, o português. Essas atividades são parte do projeto EPT-Web, que visa à tradução automática para o português de textos jornalísticos (manchetes e leads) produzidos em língua inglesa e disponíveis na versão eletrônica da primeira página do jornal norte-americano *The New York Times*. O projeto é realizado pelo NILC e tem o apoio do CNPq.

[1] http://www.linguatec.net/online/ptwebtext/index_en.shtml
[2] http://www.intertran.net
[3] http://www.systranbox.com/systran/box
[4] http://www.freetranslation.com

Índice

1. INTRODUÇÃO **3**
2. MOTIVAÇÃO PARA A AVALIAÇÃO **4**
2.1 Parâmetros de avaliação 4
3. AVALIAÇÃO LEXICAL **7**
3.1 Dicionarização 7
3.2 Ambigüidade lexical 8
4. CONCLUSÃO E PRÓXIMOS PASSOS **16**
5. BIBLIOGRAFIA CONSULTADA **17**

1. INTRODUÇÃO

O projeto EPT-Web visa ao desenvolvimento de uma ferramenta computacional para tradução de textos jornalísticos, em especial da primeira página do jornal eletrônico norte-americano The New York Times (http://www.nyt.com), de forma unidirecional (do inglês para o português) e completamente automática. Trata-se do desenvolvimento de um plug-in a ser incorporado a uma ferramenta de navegação na Web (Internet Explorer), que permita uma tradução de qualidade, em tempo real e que mantenha a formatação (diagramação, figuras, tabelas, links e outros) da primeira página do jornal eletrônico. O aplicativo representa uma tentativa de aproximação ao problema do desenvolvimento de um sistema de tradução automática de âmbito mais genérico, não restrito ao domínio dos textos a serem traduzidos. A execução do projeto adota um modelo de tradução multilingual baseado em interlíngua derivado do Projeto UNL.

O Projeto UNL tem por objetivo o desenvolvimento de um sistema que possibilite ao usuário produzir material de sua própria língua em UNL (por meio de um codificador língua natural-UNL) e disponibilizá-lo na Web, a fim de que falantes de outras línguas tenham acesso a esse material utilizando decodificadores UNL-língua natural. O Projeto inclui atualmente 15 línguas (árabe, alemão, chinês, espanhol, francês, hindi, indonésio, inglês, italiano, japonês, letão, mongol, português, russo e tailandês), cada uma atribuída a diferentes grupos de pesquisa e desenvolvimento espalhados pelo mundo e tem duração prevista de dez anos a partir de 1997. Os sistemas de codificação e decodificação estão sendo desenvolvidos por universidades, institutos de pesquisa e empresas de diversos países, o que distingue o projeto UNL como uma tentativa de conjugar esforços de especialistas em processamento automático das línguas naturais (PLN) do mundo todo. As ferramentas de PLN para o português do Brasil (UNL-Brasil) estão a cargo do Núcleo Interinstitucional de Lingüística Computacional – NILC.

Hoje, o desenvolvimento de sistemas de tradução automática exige a intervenção humana, seja na pré ou pós-edição dos textos a serem traduzidos. Isso ocorre por causa dos resultados pouco satisfatórios produzidos, o que vem conduzindo a uma revisão de metas, segundo a qual a tradução completamente automática, com qualidade, de textos arbitrários, não será atingida em curto prazo. Para a produção de resultados aceitáveis, este projeto circunscreve seu campo de ação imediato a um gênero textual bastante determinado, de dimensões reduzidas e estruturas fixas. A utilização da Universal Networking Language (UNL) - linguagem artificial para representação semântica do conteúdo ideacional dos enunciados lingüísticos - como interlíngua semântica entre o inglês e português assegura ao sistema a portabilidade necessária para sua expansão para outros domínios e gêneros (em função do caráter abrangente e generalista da UNL) e para uma plataforma multilíngüe (dado que a incorporação de novas línguas aos sistemas interlinguais pode ser feita de forma independente e modular, com a cooperação de outros grupos, também envolvidos com o desenvolvimento de sistemas para UNL).

Nesse contexto, este relatório apresenta a descrição de uma parcela das atividades realizadas durante os segundos 12 meses dos 24 previstos para o projeto, mais especificamente aquelas que envolvem o teste de ferramentas de tradução automática já disponíveis na Web: Linguatec e-translation Server (ES), Intertran (IT), Systran (SY) e FreeTranslation (FT).

2. MOTIVAÇÃO PARA A AVALIAÇÃO

A motivação para a realização destas atividades de testes é tornar possível avaliarmos, no futuro, a ferramenta desenvolvida no projeto EPTWeb – o Policarpo – de maneira comparativa, a fim de verificarmos quantos dos problemas comumente encontrados nos sistemas de tradução automática foram solucionados pela nova ferramenta e quais problemas ainda necessitam tratamento. O objetivo desta avaliação é, desta forma, reunir informações suficientes para mais tarde fazermos um estudo comparativo entre as ferramentas de Tradução Automática já existentes e o Policarpo.

Quatro ferramentas de tradução automática foram empregadas em nosso estudo: Linguatec e-translation Server (ES), Intertran (IT), Systran (SY) e FreeTranslation (FT). As quatro ferramentas permitem a passagem tanto do inglês para o português, quanto do português para o inglês, e todas estão disponíveis na Web. Os testes foram feitos apenas na tradução do inglês para o português. Para comparar as ferramentas citadas, utilizamos um corpus composto por 515 sentenças retiradas da primeira página da versão eletrônica do jornal norte-americano *The New York Times* (30 páginas recolhidas durante os dias 03 e 04 de abril de 2002). Essas 30 páginas compõem o corpus de treino do projeto EPTWeb. Temos que reconhecer que esse número é pequeno para garantir a qualidade das ferramentas analisadas, portanto deixemos claro que o nosso objetivo não é hierarquizar as ferramentas, mas discutir os problemas encontrados em suas produções.

2.1 Parâmetros de avaliação

Avaliação tem um papel de extrema importância no desenvolvimento de novas tecnologias envolvendo sistemas de processamento de língua natural (PLN). Embora todos concordem quanto às linhas gerais de avaliação, não existem, até o momento, métodos e parâmetros universalmente aceitos relacionados à avaliação de sistemas de tradução automática. Esse tipo de avaliação envolve características peculiares, nem sempre encontradas na avaliação de outros tipos de sistemas de PLN, como, por exemplo, qualidade das traduções antes da pós-edição feita por um tradutor humano, avaliando, assim, aspectos como inteligibilidade, precisão, fidelidade, emprego apropriado de estilo/registro. Em um nível mais profundo, avalia-se a facilidade em criar e atualizar dicionários, possibilidade de ampliar os pares de línguas ou domínios que compõem o sistema, e o custo-benefício em comparação com traduções humanas.

No entanto, existem três formas de avaliação que geralmente são aceitas: avaliação da adequação, avaliação diagnóstica e avaliação da performance. A avaliação da adequação é geralmente usada para estimar a adequação de sistemas de tradução automática a certos contextos operacionais. A avaliação diagnóstica objetiva identificar limitações, erros e deficiências de sistemas de tradução automática. Finalmente, a avaliação da performance pretende testar estágios do desenvolvimento do sistema.

Neste trabalho, faremos uma avaliação diagnóstica das quatro ferramentas já citadas, a fim de obter uma visão dos erros e deficiências encontrados em suas produções.

O estudo das produções das 4 ferramentas de tradução automática seguiu, primeiramente, o critério da inteligibilidade/gramaticalidade das sentenças. Assim,

dividimos as 515 sentenças em "gramaticais corretas", "gramaticais incorretas" e "agramaticais".

A noção de "gramaticalidade" e "agramaticalidade" que empregamos leva em conta a identificação das sentenças produzidas com a estrutura subjacente à língua portuguesa. Levamos em conta também a adequação das sentenças gramaticais, tanto corretas quanto incorretas, em relação à gramática normativa. Tendo isso em conta, as sentenças "gramaticais incorretas" serão aquelas que possuem estrutura frasal adequada à língua portuguesa (por exemplo, SV V SN SAdv), obedecendo às regras gramaticais sem, no entanto, obedecerem a regras semânticas. Ou seja, a sentença pode não fazer sentido algum, mas obedecendo a regras sintáticas, ela será classificada como uma sentença "gramatical incorreta". Isso mostra que sintaxe e sentido podem ser independentes entre si, como demonstrou Chomsky com a frase: *Incolores idéias verdes dormem furiosamente*.

Nessa categoria estão também as sentenças que foram traduzidas incorretamente, isto é, aquelas sentenças cujo significado não é o mesmo do original. Neste ponto, leva-se em conta alguma noção de fidelidade à sentença original.

Exemplos:

1. Trouble Ends Jordan's Season

(ES) A dificuldade de joelho termina a estação do ano de Jordão.

2. Astros Rout Brewers, 15 - 2
(IT) Torrinha Derrota Cervejeiros , 15 2

3. New York Archdiocese Turns Over List of Accused Abusers
(SY) O archdiocese de New York vira a lista de abusadores acusados

4. Sexual Abuse Lawsuits Filed Against Vatican
(FT) Processos sexuais de Abuso Arquivaram Contra Vaticano

As sentenças "gramaticais corretas" são aquelas que traduzem de uma forma aceitável o sentido da frase original. Para pertencer a esta categoria, as sentenças necessariamente precisam obedecer às regras sintáticas e semânticas da língua portuguesa. Assim, aspectos como concordância verbal e nominal são também considerados.

Exemplos:

5. Foreigners Evacuated From Bethlehem

(ES) Estrangeiros evacuados de Belém.

6. Making Books:
(IT) Fazendo Livros :

7. Technique May Improve Safety of Donated Blood
(SY) A técnica pode melhorar a segurança do sangue doado

8. Washington Hires Romar As Coach
(FT) Washington Emprega Romar Como Treinador

E, finalmente, as sentenças "agramaticais" são aquelas que não possuem nada que as identifique como uma sentença da língua portuguesa, a não ser o léxico, que, no entanto, não está ligado de forma a estruturar uma sentença pertencente ao português. Tais sentenças se apresentam como uma lista de palavras aleatoriamente conectadas. Nessa categoria encontram-se também aquelas sentenças possíveis de serem entendidas pelo usuário da ferramenta, apesar da sua "agramaticalidade". São sentenças interpretáveis, mas que precisam de pós-tratamento para serem sentenças da língua portuguesa.

Exemplos:

9. Wed Cousins Birth Defects Studied

(ES) Quarta-feira os defeitos de nascimento Cousins estudaram.

10. Executive Order Followed Energy Industry Recommendation
(IT) Ordem executivo Seguido Energia Indústria Recomendação

11. Israeli Military Sends Tanks Into Largest West Bank City
(SY) As Forças armadas Israeli Emitem Os Tanques Na Cidade A maior Do Banco Ocidental

12. Britain Scuttles Plans to Publish Iraq Dossier
(FT) A bretanha sai em disparada Planos Publicar Ficha de Iraque

Do corpus total, composto por 515 sentenças, obtivemos os seguintes números:

Tabela 1: Primeira avaliação do material gerado pelos 4 sistemas avaliados

	Linguatec e-translation Server	Intertran	Systran	FreeTranslation
Gramaticais Corretas	160 – 31,1%	73 – 14,1%	159 – 30,9%	175 – 34%
Gramaticais Incorretas	197 – 38,2%	49 – 9,6%	97 – 18,8%	93 – 18%
Agramaticais	158 – 30,7%	393 – 76,3%	259 – 50,3%	247 – 48%
TOTAL	515 – 100%	515 – 100%	515 – 100%	515 – 100%

Podemos observar na tabela 1 que nenhum sistema alcançou uma porcentagem maior que 34% de sentenças "gramaticais corretas", sendo que o sistema que apresentou o melhor desempenho foi o FreeTranslation (34%), seguido pelo Linguatec e-translation Server (31,1%) e pelo Systran (30,9%), sendo o Intertran o sistema com o pior desempenho (14,1%). Apenas o Linguatec e-translation Server apresenta-se estável em relação aos erros e acertos das sentenças geradas (31,1% de sentenças "gramaticais corretas", 38,2% de "gramaticais incorretas" e 30,7% de "agramaticais"). Isso significa que a somatória das porcentagens de sentenças "gramaticais incorretas" e "agramaticais" é bem significativa, sempre maior que 50% (66,8% para o Linguatec e-translation Server, 85,9% para o Intertran, 69,1% para o Systran e 66% para o FreeTranslation).

Tendo em conta a grande quantidade de sentenças problemáticas produzidas, fizemos uma análise dos problemas gerais que as sentenças apresentavam. Abaixo está uma lista dos itens mais recorrentes:

* dicionarização
* homonímia
* polissemia
* sinonímia
* delimitação entre palavras
* falta de letras
* concordância nominal
* concordância verbal
* regência
* colocação (posição das palavras nos grupos nominais)
* tempos verbais
* letra maiúscula
* sinais gráficos
* pontuação

Percebe-se que os problemas atingem os níveis lexical, sintático, semântico e pragmático das sentenças. Tendo em vista a grande variedade de tipos de problemas e a inter-relação que ocorre entre eles, optamos por focalizar apenas um, a fim de viabilizar um estudo mais completo. Desta forma, observando o grande número de problemas lexicais, resolvemos fazer um levantamento mais detalhado de qual tipo de problemas de léxico foram encontrados.

3. AVALIAÇÃO LEXICAL

No nível lexical, situações envolvendo dois critérios foram observadas: dicionarização e ambigüidade lexical.

3.1 Dicionarização

Este caso envolve sentenças com um ou mais termos que permanecem sem tradução. Assim, a qualidade (ou a falta) da tradução advém de quão correta e compreensivelmente os termos foram dicionarizados. Os problemas mais freqüentes ocorrem principalmente com nomes próprios ou palavras derivadas delas, principalmente topônimos, embora ocorram também casos em que substantivos comuns não são traduzidos. Além dos casos em que as palavras são mantidas na sua forma original (no caso, em inglês), existem também ocorrências em que palavras inexistentes no léxico da língua portuguesa foram geradas.

Exemplos:

13. Israelis Battle Gunmen in Nablus

(ES) As israelitas combatem **Gunmen** em Nablus.

14. Bombers Gloating in Gaza
(IT) Bombardeiros **Gloating** em Gaza

15. Boston Marathon Increases Security
(SY) O **Marathon** De Boston Aumenta A Segurança

16. Farms and Growth Threaten a Sea and Its Creatures
(FT) As fazendas e Crescimento Ameaçam um Mar e **Seuas** Criaturas

3.2 Ambigüidade lexical

O problema de ambigüidade lexical revela que a dicionarização sozinha não é suficiente para solucionar questões envolvendo léxico. É também necessário desenvolver estratégias de seleção entre as várias acepções dicionarizadas. A questão da ambigüidade lexical na tradução possui as seguintes formas de classificação: ambigüidade categorial, que ocorre quando a palavra da língua alvo apresenta classes gramaticais diferentes (em relação às opções dadas na tradução da língua fonte), ou ambigüidade de sentido, que ocorre quando a palavra da língua fonte possui várias opções de tradução dentro de uma mesma classe gramatical. Desta forma, a presença freqüente de homônimos, polissemias e sinonímias, tanto em português quanto em inglês, faz com que seja necessário o uso de estratégias de desambigüização.

Observe que o problema, quando se trata de escolha entre as várias opções dentro de uma mesma classe de palavras, é que a ferramenta calcula, corretamente, que se trata de um substantivo, mas, entre os muitos substantivos candidatos, ela termina por selecionar um substantivo que não cabe no contexto. Nesta avaliação, foi feito um levantamento de todas as ocorrências de termos ambíguos, sem categorizar as ambigüidades em ambigüidade categorial ou de sentido. Da mesma forma, não houve subcategorização da ambigüidade de sentido, o que permitiria dividirmos os termos ambíguos de cada categoria em homonímia, polissemia e sinonímia. Portanto, nesta primeira avaliação, levamos em conta todas as ocorrências de termos ambíguos.

Exemplos:

*17. Israeli Military Sends **Tanks** Into Largest West Bank City*
(ES) O exército israelense envia **reservatórios** em cidade West Bank mais grande.

*18. **Bush** in Predicament Over Mideast*
(IT) **Arbusto** em Impasse Sobre Mideast

*19. AOL Makes Case in Microsoft **Trial***
(SY) AOL faz o caso na **experimentação** de Microsoft

*20. Embalming Class Used Body Without Permission, **Suit** Says*
(FT) A Classe de Embalming Usou Corpo Sem Permissão, **Terno** Diz

A seguir estão as tabelas 2, 3, 4 e 5, com os números relacionados às análises feitas levando em conta a quantidade de sentenças que apresentaram o problema específico de ambigüidade lexical. Esses são os números de sentenças que apresentaram o problema. Deve-se ficar atento ao fato de que em muitos casos, mais de um termo lexicalmente ambíguo traduzido erroneamente foi encontrado em uma mesma sentença.

Tabela 2: Sentenças com palavras lexicalmente ambíguas – Linguatec e-translation Server

	Sentenças com problemas de ambigüidade lexical	% em relação o total de sentenças com problemas	% em relação ao total de sentenças por categoria	% em relação ao número total de sentenças (515)
Gramaticais Corretas	20	7,20%	12,50%	3,90%
Gramaticais Incorretas	145	51,90%	73,60%	28,20%
Agramaticais	114	40,90%	72,20%	22,10%
Total	**279**	**100,00%**	***	**54,10%**

Tabela 3: Sentenças com palavras lexicalmente ambíguas – Intertran

	Sentenças com problemas de ambigüidade lexical	% em relação o total de sentenças com problemas	% em relação ao total de sentenças por categoria	% em relação ao número total de sentenças (515)
Gramaticais Corretas	03	0,80%	4,10%	0,60%
Gramaticais Incorretas	29	8,00%	59,20%	5,60%
Agramaticais	329	91,20%	83,70%	63,90%
Total	**361**	**100,00%**	***	**70,10%**

Tabela 4: Sentenças com palavras lexicalmente ambíguas – Systran

	Sentenças com problemas de ambigüidade lexical	% em relação o total de sentenças com problemas	% em relação ao total de sentenças por categoria	% em relação ao número total de sentenças (515)
Gramaticais Corretas	24	8,80%	15,10%	4,70%
Gramaticais Incorretas	78	28,70%	80,40%	15,10%
Agramaticais	170	62,5%	65,60%	33,00%
Total	**272**	**100,00%**	***	**52,80%**

Tabela 5: Sentenças com palavras lexicalmente ambíguas – FreeTranslation

	Sentenças com problemas de ambigüidade lexical	% em relação o total de sentenças com problemas	% em relação ao total de sentenças por categoria	% em relação ao número total de sentenças (515)
Gramaticais Corretas	26	9,60%	14,80%	5,00%
Gramaticais Incorretas	69	25,50%	74,20%	13,40%
Agramaticais	176	64,90%	71,30%	34,20%
Total	**271**	**100,00%**	***	**52,60%**

Podemos observar que os quatro sistemas apresentaram uma porcentagem maior do que 50% de sentenças com problemas específicos de ambigüidade lexical, sendo que o sistema que apresentou a porcentagem maior foi o Intertran (70,10%), e o que apresentou a menor porcentagem foi o FreeTranslation (52,60%). Vale especificar que essas sentenças podem apresentar outros problemas, como, por exemplo, concordância verbal/nominal, não tradução, regência, colocação, etc., mas todas apresentaram um ou mais termos com problemas de ambigüidade lexical.

Observando a porcentagem em relação ao total de sentenças com problemas de ambigüidade lexical, com exceção do Linguatec e-translator Server, todas as ferramentas concentram a maior parte de sentenças problemáticas na categoria "agramaticais". Em relação à porcentagem, considerando-se o total de sentenças por categorias, as ferramentas Linguatec e-translation Server, Systran e Free Translation concentraram os problemas de ambigüidade lexical na categoria "gramaticais incorretas", enquanto a Intertran concentra a maior parte dos problemas na categoria "agramaticais".

A partir da primeira observação, podemos deduzir que ainda que os problemas de ambigüidade lexical fossem resolvidos, as sentenças ainda permaneceriam incorretas,

uma vez que, pertencendo à categoria "agramatical", a ambigüidade se apresenta como apenas um dos problemas da sentença. Por sua vez, o fato da ferramenta Linguatec e-translation Server, Systran e FreeTranslation concentrarem, em números relativos, a maior parte das sentenças com problemas de ambigüidade lexical na categoria "gramaticais incorretas", sugere que, uma vez resolvido esse problema semântico, as sentenças poderiam, em sua maioria, ser categorizadas como "gramaticais corretas". Só para exemplificar, no caso de todas as sentenças "gramaticais incorretas" com problemas de ambigüidade lexical apresentadas pela ferramenta Free Translation (69 sentenças, de um total de 93, o que corresponde a 64,20%) terem seus problemas solucionados, isso aumentaria o número de sentenças satisfatórias produzidas pela ferramenta de 175 (34%) para 268 (52%). Vale dizer que essa estimativa está sendo feita considerando que o único problema dessa sentença seja a ambigüidade lexical, uma vez que elas já apresentam a estrutura sintática da língua portuguesa. De qualquer forma, fica claro com esses números que a ambigüidade lexical compromete profundamente a qualidade das traduções produzidas automaticamente.

Um segundo critério de avaliação foi a subdivisão das ambigüidades lexicais encontradas de acordo com as categorias lexicais: substantivo, adjetivo, verbo, conjunção, preposição, advérbio, pronome, numeral, nome próprio e prefixo. Vale dizer que a categoria "nome próprio" são aqueles termos que designam nomes de times, de pessoas, de lugares e de empresas os quais não têm tradução para o português e que, portanto, devem permanecer no inglês. Abaixo estão os números para cada uma das três categorias: "gramaticais corretas", "gramaticais incorretas" e "agramaticais", para cada um dos quatro sistemas avaliados.

Tabela 6: Categoria das ambigüidades lexicais
Sentenças gramaticais corretas / Linguatec e-translation Server

Substantivo	08	40,00%
Verbo	06	30,00%
Adjetivo	04	20,00%
Nome Próprio	01	5,00%
Pronome	01	5,00%
TOTAL	**20**	**100,00%**

Tabela 7: Categoria das ambigüidades lexicais – Sentenças gramaticais corretas / Intertran

Substantivo	02	100,00%
TOTAL	**02**	**100,00%**

Tabela 8: Categoria das ambigüidades lexicais – Sentenças gramaticais corretas / Systran

Substantivo	11	57,20%
Verbo	5	19%
Nome próprio	3	14,30%
Adjetivo	2	9,50%
TOTAL	**21**	**100,00%**

Tabela 9: Categoria das ambigüidades lexicais – Sentenças gramaticais corretas / FreeTranslation

Substantivo	15	50,00%
Nome próprio	9	30%
Verbo	4	13,30%
Adjetivo	2	6,70%
TOTAL	**30**	**100,00%**

Um número bem pequeno de termos ambíguos foi encontrado nessa categoria, uma vez que o critério para classificar as sentenças como "gramaticais corretas" era exatamente estar o mais próximo possível de uma tradução humana. Mas ainda assim, foram encontradas algumas ocorrências de ambigüidades lexicais que comprometiam a qualidade da tradução. Para os quatro sistemas, a maioria das ocorrências foi com substantivos, sendo que para a ferramenta Intertran, apenas substantivos apresentaram problema de ambigüidade. Abaixo estão alguns exemplos:

21. Restaurants Accused of ***Tax Evasion***
(ES) Restaurantes acusados de **evasão fiscal**.

22. After decades of false starts, interactive television is finally arriving in ***living rooms***.
(SY) Após décadas de começos falsos, a televisão interativa está chegando finalmente em **quartos vivos**.

23. Contrary to widely held beliefs, first cousins can have ***children*** *together without a great risk of birth defects or genetic disease.*
(FT) Contrário a crenças largamente seguradas, primeiros prioes pode ter **crianças** juntas sem grande risco de defeitos de nascimento ou doença genética.

Podemos perceber que a qualidade foi prejudicada pelo problema de ambigüidade dos termos "tax evasion", "living rooms" e "children". Nos primeiro, segundo e quarto casos, o problema é sutil: tanto "tax evasion" pode ser traduzido por "evasão fiscal" quanto "children" por "crianças". No entanto, as traduções mais apropriadas seriam: "sonegação de imposto", "cinema" e "filhos" respectivamente. O problema seria mais estilístico. No terceiro exemplo, o sistema não reconhece "living rooms" como uma única palavra (salas), e acaba traduzindo termo a termo uma palavra composta.
Abaixo estão alguns exemplos de sentenças "gramaticais corretas" com problema de ambigüidade lexical em verbos, adjetivos e nomes próprios respectivamente:

24. Fighting ***Engulfs*** *Nablus, EU Sends Peace Mission*
(ES) A luta **traga** Nablus, EU envia missão de paz.

25. Study: Birth Defect Risk ***Lower***
(FT) O estudo: **Abaixamento** de Risco de Defeito de Nascimento

26. Knee Problem Ends ***Jordan****'s Season*
(SY) O Problema Do Joelho Termina Estação De **Jordão**

As tabelas 10, 11, 12 e 13 mostram os números relacionados à categoria "gramaticais incorretas". Nesta categoria, o problema de ambigüidade lexical é, na maioria das vezes, o responsável pela falta de adequação da sentença.

Tabela 10: Categoria das ambigüidades lexicais
Sentenças gramaticais incorretas / Linguatec e-translation Server

Substantivo	64	45,40%
Verbo	41	29,10%
Nome Próprio	23	16,30%
Adjetivo	11	7,80%
Conjunção	01	0,70%
Prefixo	01	0,70%
TOTAL	**141**	**100,00%**

Tabela 11: Categoria das ambigüidades lexicais – Sentenças gramaticais incorretas / Intertran

Substantivo	11	61,10%
Verbo	04	22,20%
Nome Próprio	03	16,70%
TOTAL	**18**	**100,00%**

Tabela 12: Categoria das ambigüidades lexicais – Sentenças gramaticais incorretas / Systran

Substantivo	35	39,30%
Verbo	25	28,10%
Nome próprio	19	21,40%
Adjetivo	9	10,10%
Advérbio	1	1,10%
TOTAL	**89**	**100,00%**

Tabela 13: Categoria das ambigüidades lexicais – Sentenças gramaticais incorretas / FreeTranslation

Substantivo	42	53,20%
Verbo	16	20,20%
Nome próprio	16	20,20%
Adjetivo	3	3,80%
Advérbio	1	1,30%
Pronome	1	1,30%
TOTAL	**79**	**100,00%**

O ranking de categorias lexicais, nas sentenças "gramaticais incorretas", para os sistemas Systran e Free Translation é igual: substantivo, verbo, nome próprio, adjetivo e advérbio, sendo que o FreeTranslation apresenta a ocorrência de 1 pronome. O Linguatec e-translation Server apresenta essa mesma hierarquização com a diferença de apresentar uma conjunção e um prefixo e não apresentar advérbios. Já o Intertran apresenta apenas problemas com substantivos, verbos e nomes próprios. Abaixo estão alguns exemplos de sentenças "gramaticais incorretas" que apresentam problemas na tradução de substantivos (exemplos 27 e 28), verbos (exemplos 29 e 30) e nomes próprios (exemplos 31 e 32), respectivamente:

*27. (See the **cartoon**.)*
(ES) (Veja a **caricatura**.).

*28. Living on Internet Time, in Another **Age***
(SY) Vivendo no tempo do Internet, em uma outra **idade**

*29. With an Organic Sensor, a Food Wrapper **Sniffs Out** Trouble*
(ES) Com um sensor orgânico, um envoltório de comida **funga fora de** dificuldade.

*30. Bush **Sending** Powell to Middle East*
(SY) Bush que **emite** Powell ao leste médio

*31. Firestone Faces Trial on Tire **Rollover***
(ES) O Firestone encara prova sobre pneu **renovável**

*32. **Muddy Waters**, American blues musician, was born.*
(FT) **Águas lamacentas**, músico Americano de blues, nascia.

E, finalmente, abaixo estão as tabelas com os números relacionados à categoria "agramaticais", que, entre muitos problemas (colocação, estrutura, concordância), apresentou um número elevado de sentenças com problemas de ambigüidade lexical. Podemos observar um número alto de palavras com problemas de ambigüidade (135 termos para o Linguatec e-translation Server, 324 termos para o Intertran, 221 termos para o Systran e 225 termos para o FreeTranslation).

Tabela 14: Categoria das ambigüidades lexicais
Sentenças agramaticais / Linguatec e-translation Server

Verbo	43	31,90%
Substantivo	38	28,10%
Nome Próprio	29	21,50%
Adjetivo	19	14,10%
Advérbio	4	3,00%
Preposição	1	0,70%
Pronome	1	0,70%
Total	**135**	**100,00%**

Tabela 15: Categoria das ambigüidades lexicais – Sentenças agramaticais / Intertran

Substantivo	120	37,00%
Verbo	107	33,00%
Nome Próprio	52	16,10%
Adjetivo	36	11,10%
Advérbio	6	1,90%
Pronome	1	0,30%
Prefixo	1	0,30%
Numeral	1	0,30%
TOTAL	**324**	**100%**

Tabela 16: Categoria das ambigüidades lexicais – Sentenças agramaticais/ Systran

Substantivo	85	38,50%
Verbo	73	33,00%
Nome próprio	39	17,70%
Adjetivo	18	8,10%
Preposição	4	1,80%
Expressão	2	0,90%
TOTAL	**221**	**100,00%**

Tabela 17: Categoria das ambigüidades lexicais – Sentenças agramaticais/ FreeTranslation

Substantivo	84	37,30%
Verbo	54	24,00%
Nome próprio	56	24,90%
Adjetivo	25	11,10%
Preposição	2	0,90%
Advérbio	2	0,90%
Expressão	2	0,90%
TOTAL	**225**	**100,00%**

Podemos observar que, novamente, o ranking das categorias lexicais é o mesmo para os sistemas Systran, FreeTranslation e Intertran: substantivo, verbo, nome próprio e adjetivo abarcam a maioria absoluta das categorias lexicais com problema de ambigüidade: 97,3% para os dois primeiros e 95,6% para o terceiro. Abaixo estão alguns exemplos de termos problemáticos nas categorias substantivo (exemplos 33, 34 e 35), verbo (exemplos 36, 37 e 38), nome próprio (exemplos 39 e 40) e adjetivo (exemplos 41 e 42), respectivamente, para os três sistemas:

*33. Peck Is New **Coach** of Lady Gators*
(SY) O peck é **ônibus** novo da senhora Gators

*34. Condom as a Problem Word: Iran Grapples With a Surge in **AIDS***
(FT) O preservativo como uma Palavra de Problema: Irã Segura Com um Oscilar em **AJUDAS**

*35. AP Corrects Burger Beef **Story***
(IT) AP Correto Broto Bife **Conto**

*36. **Click** Here to Receive 50% Off Home Delivery of The New York Times Newspaper.*
(SY) **Estale** aqui para receber 50% fora da entrega home do jornal dos tempos de New York.

*37. **Check** them **out**, or post any wine-related topics.*
(FT) **Verifique**-os **para fora**, ou coloca qualquer temas vinho-relacionados.

*38. Kirch in Danger of Bankruptcy After Rescue Talks **Break Down***
(IT) Kirch em Perigo de Falência Após Salvamento Falas **Brecha Abaixo**

*39. **Knight Trading** Sees First - Quarter Loss*
(SY) **Negociar Do Knight** Vê Primeiramente - A Perda De um quarto

*40. **Bush** to Make Statement on Mideast*
(FT) A **mata** Fazer Declaração em Meio oriente

*41. In **Free** Music Software, a Hidden Fee-Based Service*
(SY) No software livre da música, um serviço Taxa-Baseado escondido

*42. Hundreds Arrested in **Afghan** Plot*
(IT) Centenas Detido em **Xale** Intriga

Já o Linguatec e-translation Server apresenta uma diferença em relação aos outros sistemas: a maioria dos problemas, com 31,9% das ocorrências são verbos, depois aparecem os substantivos (28,1%), seguidos dos nomes próprios (21,5%) e adjetivos (14,1%). Apresenta ainda algumas ocorrências de advérbios, preposições e pronomes. Alguns exemplos de sentenças com verbos, substantivos e nomes próprios para essa ferramenta estão abaixo:

*43. Pirates 5, Mets 3: New-Look Mets Have Kinks to **Iron Out***
(ES) Piratas 5, Mets 3: Mets New-Look tem voltas de **engomar**.

*44. Governor's Election **Arm** Indicted in Illinois Scandal*
(ES) O **braço** de eleição de Governor indiciou em escândalo Illinois.

*45. Webber Leads **Kings** to Win Over **Pistons***
(ES) Webber porta **rei** para convencer **êmbolos**.

Além dessas observações relacionadas pontualmente às categorias "gramaticais corretas", "gramaticais incorretas" e "agramaticais", outras observações podem ser feitas relacionando os três conjuntos de tabelas, visando o conjunto geral das categorias lexicais que apresentaram problemas de ambigüidades lexicais. Abaixo estão as tabelas com os números e porcentagens totais para os sistemas:

Tabela 18: Categoria das ambigüidades lexicais – Total / Linguatec e-translation Server

Substantivo	106	36,70%
Verbo	86	29,80%
Nome próprio	54	18,70%
Adjetivo	34	11,80%
Advérbio	4	1,40%
Pronome	2	0,70%
Preposição	1	0,30%
Conjunção	1	0,30%
Prefixo	1	0,30%
TOTAL	**289**	**100,00%**

Tabela 19: Categoria das ambigüidades lexicais – Total / Intertran

Substantivo	133	38,70%
Verbo	111	32,30%
Nome próprio	55	15,90%
Adjetivo	36	10,50%
Advérbio	6	1,70%
Pronome	1	0,30%
Prefixo	1	0,30%
Numeral	1	0,30%
TOTAL	**344**	**100,00%**

Tabela 20: Categoria das ambigüidades lexicais – Total / FreeTranslation

Substantivo	117	40,00%
Verbo	92	31,40%
Nome próprio	51	17,40%
Adjetivo	25	8,50%
Preposição	2	0,70%
Advérbio	1	0,30%
Pronome	0	0,00%
Sem classificação	5	1,70%
TOTAL	**293**	**100,00%**

Tabela 21: Categoria das ambigüidades lexicais – Total / Systran

Substantivo	116	39,60%
Verbo	71	24,10%
Nome próprio	70	23,80%
Adjetivo	28	9,50%
Advérbio	3	1,00%
Preposição	1	0,30%
Pronome	1	0,30%
Sem classificação	4	1,40%
TOTAL	**294**	**100,00%**

Podemos observar que, considerando as ocorrências em geral, as categorias lexicais substantivo, verbo, nome próprio e adjetivo computam a maioria absoluta de termos com problema de ambigüidade: 40% para o Systran, 39,60% para o FreeTranslation, 38,8% para o Intertran e 36,40% para o Linguatec e-translation Server.

4. CONCLUSÃO E PRÓXIMOS PASSOS

Os testes feitos são parte do projeto EPTWeb e tornará possível a avaliação da ferramenta de tradução automática (Policarpo) desenvolvida por ele.

Após uma avaliação diagnóstica das ferramentas de Tradução Automática: Linguatec e-translation Server (ES), Intertran (IT), Systran (SY) e FreeTranslation (FT), pudemos observar que nenhuma das ferramentas alcançou um número satisfatório de

sentenças consideradas "gramaticais corretas", ou seja, sentenças que traduzem de uma forma aceitável o sentido da frase original, isso porque nas quatro ferramentas a somatória das sentenças "gramaticais incorretas" (aquelas que obedecem regras gramaticais, mas não obedecem regras semânticas) e "agramaticais" (não possuem nada que as identifique como uma sentença da língua portuguesa) sempre ultrapassa 50% do número total de sentenças do corpus.

Levando em conta os dados levantados na avaliação realizada, podemos afirmar que o número de ambigüidades reflete a qualidade da ferramenta e a solução dessas questões envolvendo ambigüidade lexical mostra-se como um dos caminhos para a obtenção de resultados mais satisfatórios nas produções das ferramentas de traduções automáticas.

Depois de verificados os problemas mais comuns, poderemos constatar quais deles a ferramenta Policarpo consegue suprir e quais ainda precisam de tratamento. Entre as próximas atividades, portanto, estão a análise mais detalhada das ambigüidades lexicais encontradas e a avaliação dos resultados apresentados pelo Policarpo.

5. REFERÊNCIAS

Cole, R. A.; Mariani, J.; Uszkoreit, H.; Zaenen, A.; Zue, V. (1996). *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge.

Linguatec Language Technologies. **Linguatec e-translation Server**. Disponível em: <http://www.linguatec.net/online/ptwebtext/index_en.shtml>. Acesso em: 19 maio 2004.

Nunes M.G.V.; Martins, R.T.; Fossey, M.F.; Pedrolongo, T. (2002). A construção do corpus e dos dicionários Inglês-UNL e UNL-português para o projeto EPT-Web. Série de Relatórios do NILC, NILC-TR-02-24.

Oliveira Jr. O.N.; Martins, M.S.; Marchi, A.R.; Martins, R.T. (2000) A critical analysis of the performance of English-Portuguese-English MT systems. Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'2000), pp.85-92. Atibaia, SP.

SDL International e-nabling global business. **FreeTranslation**. Disponível em: <http://www.freetranslation.com>. Acesso em: 19 maio 2004.

Systran Information and Translation Technologies. **Systran**. Disponível em: <http://www.systranbox.com/systran/box>. Acesso em: 19 maio 2004.

Translation Experts.com. **Intertran**. Disponível em: <http://www.intertran.net>. Acesso em: 19 maio 2004.