Novas Ferramentas e Recursos Linguísticos para a Tradução Automática

Por ocasião d'O Fim do Início de uma Nova Era no Processamento da Língua Portuguesa

Anabela Barreiro
Faculdade de Letras da Universidade do Porto & CLUP-Linguateca
New York University
barreiro_anabela@hotmail.com

1 Enquadramento

A comemoração dos 10 Anos da Linguateca é uma data que marca o início de uma nova fase na história do processamento da língua portuguesa. Ao longo de 10 anos a Linguateca teve um papel relevante na formação de recursos humanos, nomeadamente de linguistas computacionais e revelou um compromisso e empenhamento assumidos na ampliação do número e variedade de recursos linguísticos e ferramentas partilháveis disponíveis em domínio público: mais de uma dezena de novas ferramentas públicas ao serviço da comunidade, três grandes projectos de avaliação a nível nacional, várias participações em projectos internacionais de avaliação, centenas de publicações nas mais variadas conferências e revistas da especialidade, são os resultados de uma equipa que trabalhou muito em pouco tempo.

Apesar de o meu estatuto ser, na sua globalidade, oficial e logisticamente externo à Linguateca, o meu espírito rege-se pelos ideais defendidos pela Linguateca. Quando em 1998 em Portugal se discutia o futuro do processamento computacional do português, encontrava-me a trabalhar numa empresa de tradução automática nos Estados Unidos. Tive oportunidade de vir a Lisboa aquando do debate público em torno do Livro Branco e desde então sempre acompanhei de perto e atentamente as actividades que se foram desenvolvendo ao longo destes anos, participando, ainda que de forma breve em algumas das suas actividades, nomeadamente na criação de um programa de estágio em tradução automática, na anotação da Floresta Sintá(c)tica [1] [2], na organização das Morfolimpíadas e construção/revisão da lista dourada [3] [4], e no desenvolvimento de ferramentas para avaliação da tradução automática [5] [6], entre outras. Em 2003, a Linguateca iniciou a divulgação e dinâmica de actividades de avaliação conjuntas em várias áreas do processamento de língua natural, mobilizando a comunidade para a criação de um grupo de avaliação de tradução automática, o grupo ARTUR, apresentado no AVALON 2003. Em colaboração com a Universidade do Porto, trabalhou-se no desenvolvimento de uma ferramenta automática de geração de baterias de teste [7] e um programa de categorização de erros brevemente descritos em [8]. Desde essa data, a minha ligação à Linguateca estreitou-se com o meu doutoramento a ser co-orientado pela Professora Belinda Maia, responsável pelo pólo do Porto. Desde então, mais conhecimentos acerca dos problemas associados com a preservação de significado no processo de tradução foram adquiridos, incluindo os problemas levantados por idiomas, coloquialismos, usos metafóricos, entre outros. Na etapa final deste projecto, posso testemunhar o papel dos recursos da Linguateca na minha tese e os resultados que obtive e que gostaria de apresentar e retribuir à comunidade linguística. É no espírito, ideais e prática da Linguateca, a visão da língua como um bem comum e a partilha de conhecimento e recursos para o avanço do processamento da língua portuguesa que o trabalho que a seguir é apresentado se enquadra.

2 Tradução Automática com Conhecimento Linguístico Parafrástico

O projecto de doutoramento que se apresenta consiste no melhoramento da tradução automática através de um conhecimento estritamente linguístico sobre paráfrases. Neste trabalho, os corpora anotados disponibilizados pela Linguateca, nomeadamente os corpora anotados do COMPARA [9], serviram como ponto de partida para a inventariação de fenómenos linguísticos e de criação de algumas regras parafrásticas bilingues. Posteriormente, foram desenvolvidos recursos para a tradução automática de português para inglês com base em recursos do sistema OpenLogos. A parte monolingue desses recursos deu origem ao Port4NooJ, um sistema baseado em ontologias lexicais, descrito em [10] e disponível publicamente em [11] e [12]. O Port4NooJ foi construído com base em dicionários e gramáticas locais, com conteúdo sintáctico e semântico, criados no ambiente de desenvolvimento linguístico NooJ [13]. Os recursos linguísticos criados para o Port4NooJ já foram integrados no Corpógrafo [14] [15] e estão a ser utilizados na criação de novos recursos derivados, nomeadamente de um dicionário de expressões multipalavra e de duas ferramentas automáticas que permitem gerar e reutilizar esses recursos, ambos geradores de paráfrases, o ReWriter e o ParaMT, que apresentamos a seguir. Ambos os parafraseadores integram a aplicação de conhecimento da estrutura argumental de predicados [16] [17]. A análise sintáctico-semântica é feita no âmbito do quadro teórico do léxico-gramática [18] [19], que assenta nos

princípios da gramática transformacional harrissiana [20] [21]. Para ilustrar o funcionamento dos parafraseadores são seleccionadas paráfrases de construções com verbos suporte elementares, tais como "*fazer uma visita a*", que podem ser parafraseados por exemplo por verbos lexicais semanticamente fortes, tais como "*visitar*" ou variantes estilísticas desses verbos (verbos suporte não elementares), tais como "*efectuar uma visita a*", entre outras. As construções com verbos suporte têm sido estudadas de modo extensivo tanto do ponto de vista teórico como prático em várias línguas incluindo o português [22] [23] e [24] e como tal apresentam-se como um ponto de partida sólido para o parafraseamento.

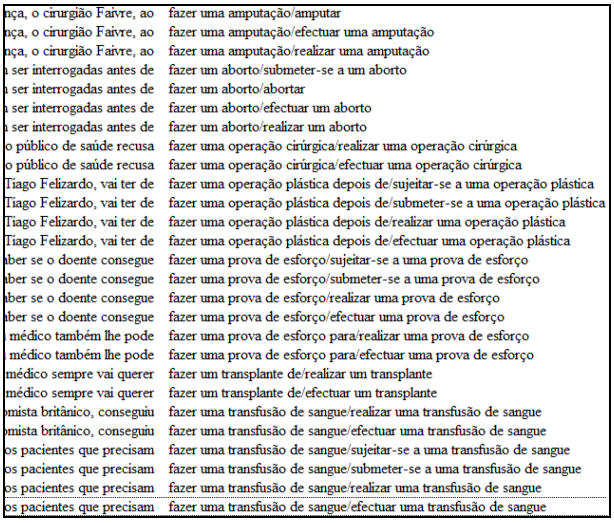
2.1 ReWriter: um Parafraseador Monolingue

O ReWriter é um parafraseador multifuncional autónomo, usado para a geração de paráfrases monolingues, com aplicações, entre outras, na preparação de textos e escrita de linguagem controlada, nomeadamente na pré-edição de texto para a tradução automática, e em funcionalidades alargadas na pesquisa e extracção de informação mais sofisticada do ponto de vista linguístico. No seu estado actual, o ReWriter reconhece e extrai construções com verbos suporte elementares a partir de textos, tais como *fazer uma operação*. Seguidamente, através de uma gramática local, a construção com verbo suporte pode ser mapeada a um verbo lexical correspondente ou a uma construção com verbo suporte não elementar que lhe seja equivalente, tal como *operar* ou *realizar uma operação*. A informação acerca das associações semânticas e morfossintácticas entre estes equivalentes encontram-se no dicionário, no caso de construções mais variáveis ou na base de dados parafrástica, no caso de construções mais cristalizadas. A mesma gramática pode conter um comando de reescrita que permite converter estas construções nos seus equivalentes parafrásticos. O funcionamento do ReWriter pode ser interactivo ou automático. Em modo interactivo, o ReWriter pode ser integrado em processadores de texto da mesma forma que os sinónimos são já aplicados. O utilizador pode sublinhar a construção com verbo suporte e clicar com o botão da direita do rato para ver quais as paráfrases que são sugeridas para essa construção, permitindo ao utilizador escolher aquela que for mais adequada ao contexto particular no texto em fase de edição. Em modo automático, a substituição é feita em simultâneo, i.e., a construção com verbo suporte é automaticamente convertida para um verbo lexical ou para outra paráfrase que esteja classificada com o valor mais alto no índice parafrástico para aquela construção, como por exemplo, uma construção com um verbo suporte não elementar de acordo com o estilo que o utilizador escolha. A paráfrase mais adequada para aquela construção será disparada automaticamente. Tanto o dicionário, como a base de dados parafrástica contém apenas lemas de construções com verbos suporte. As formas flexionadas são obtidas através do sistema flexional do Port4NooJ. Pretende-se gerar futuramente paráfrases de outros tipos de expressões e oferecer um leque vasto de alternativas que o utilizador possa utilizar de acordo com o estilo que pretenda para o seu texto.
A Fig. 1 mostra uma concordância onde algumas construções com verbos suporte são reconhecidas e parafraseadas como verbos lexicais.

gosto de ver o comboio a	fazer corridas /correr	à velocidade máxima ao lon
o de cheque especial para	fazer doações /doar	às entidades que escolher. A
res e, quando é preciso ir	fazer filmagens/filmar	fora do estúdio, às vezes fic
e queria trocar de pares e	fazer um jogo /jogar	ao melhor de três sets , mas
dra deu-me um papel para	fazer uma lista de/listar	todas as coisas boas que ex
res foram à caracterização	fazer uns retoques/retocar	, outros estão a descansar n

Fig. 1: Reconhecimento e parafraseamento monolingue de construções com verbos suporte (construção com verbo suporte / verbo lexical equivalente)

A Fig. 2 mostra uma concordância onde construções com verbos suporte que co-ocorrem com nomes predicativos ligados à área biomédica, tais como *fazer uma operação*, são reconhecidas e parafraseadas com verbos lexicais, tais como *operar*, ou variantes estilísticas léxico-sintácticas (verbos suportes não elementares) das construções com verbos suporte originais, tais como *realizar uma operação* ou *submeter-se a uma operação*. Conhecimento acerca da estrutura argumental do predicado permite a distinção de diferentes variantes estilísticas. Por exemplo, as variantes estilísticas *sujeitar-se a* e *submeter-se a* são apenas utilizadas nos casos em que o sujeito é um paciente.

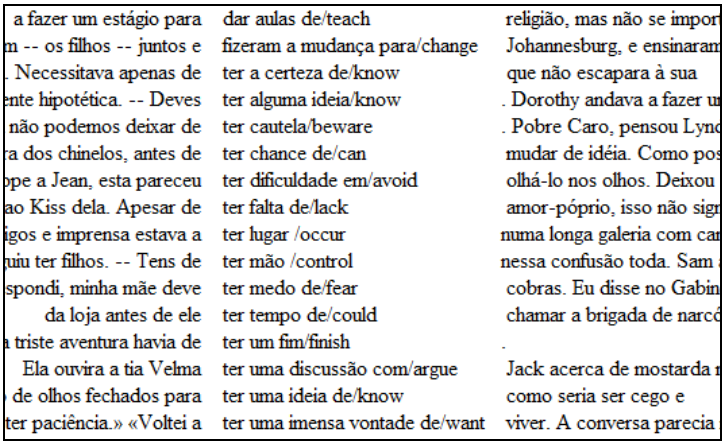


nça, o cirurgião Faivre, ao	fazer uma amputação/amputar
nça, o cirurgião Faivre, ao	fazer uma amputação/efectuar uma amputação
nça, o cirurgião Faivre, ao	fazer uma amputação/realizar uma amputação
a ser interrogadas antes de	fazer um aborto/submeter-se a um aborto
a ser interrogadas antes de	fazer um aborto/abortar
a ser interrogadas antes de	fazer um aborto/efectuar um aborto
a ser interrogadas antes de	fazer um aborto/realizar um aborto
o público de saúde recusa	fazer uma operação cirúrgica/realizar uma operação cirúrgica
o público de saúde recusa	fazer uma operação cirúrgica/efectuar uma operação cirúrgica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/sujeitar-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/submeter-se a uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/realizar uma operação plástica
Tiago Felizardo, vai ter de	fazer uma operação plástica depois de/efectuar uma operação plástica
aber se o doente consegue	fazer uma prova de esforço/sujeitar-se a uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/submeter-se a uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/realizar uma prova de esforço
aber se o doente consegue	fazer uma prova de esforço/efectuar uma prova de esforço
médico também lhe pode	fazer uma prova de esforço para/realizar uma prova de esforço
médico também lhe pode	fazer uma prova de esforço para/efectuar uma prova de esforço
médico sempre vai querer	fazer um transplante de/realizar um transplante
médico sempre vai querer	fazer um transplante de/efectuar um transplante
omista britânico, conseguiu	fazer uma transfusão de sangue/realizar uma transfusão de sangue
omista britânico, conseguiu	fazer uma transfusão de sangue/efectuar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/sujeitar-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/submeter-se a uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/realizar uma transfusão de sangue
os pacientes que precisam	fazer uma transfusão de sangue/efectuar uma transfusão de sangue

Fig. 2: Reconhecimento e parafraseamento de construções com verbos suporte que co-ocorrem com nomes predicativos da área biomédica (construção com verbo suporte / verbo lexical equivalente ou variante estilística)

2.2 ParaMT: um Parafraseador Bilingue/Multilingue

O ParaMT é um parafraseador bilingue/multilingue que opera como uma função integrada em sistemas de tradução automática do MTLab[1] e é utilizado essencialmente para a geração de equivalentes de tradução [25]. O processo de reconhecimento de uma construção com verbo suporte em texto é idêntico à do ReWriter. As gramáticas locais instruem o programa a converter automaticamente a expressão da língua de partida num equivalente parafrástico na língua de chegada. Esse equivalente pode ser um verbo lexical ou uma variante estilística. A Fig. 3 mostra o parafraseamento de uma construção com verbo suporte em português num verbo lexical em inglês.



a fazer um estágio para	dar aulas de/teach	religião, mas não se impor
m -- os filhos -- juntos e	fizeram a mudança para/change	Johannesburg, e ensinaram
. Necessitava apenas de	ter a certeza de/know	que não escapara à sua
ente hipotética. -- Deves	ter alguma ideia/know	. Dorothy andava a fazer u
não podemos deixar de	ter cautela/beware	. Pobre Caro, pensou Lyn
a dos chinelos, antes de	ter chance de/can	mudar de idéia. Como pos
ope a Jean, esta pareceu	ter dificuldade em/avoid	olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de	ter falta de/lack	amor-próprio, isso não sig
igos e imprensa estava a	ter lugar /occur	numa longa galeria com ca
uiu ter filhos. -- Tens de	ter mão /control	nessa confusão toda. Sam
spondi, minha mãe deve	ter medo de/fear	cobras. Eu disse no Gabin
da loja antes de ele	ter tempo de/could	chamar a brigada de narc
a triste aventura havia de	ter um fim/finish	.
Ela ouvira a tia Velma	ter uma discussão com/argue	Jack acerca de mostarda
de olhos fechados para	ter uma ideia de/know	como seria ser cego e
ter paciência.» «Voltei a	ter uma imensa vontade de/want	viver. A conversa parecia

Fig. 3: Reconhecimento e parafraseamento bilingue de construções com verbos suporte (construção com verbo suporte em português / verbo lexical equivalente em inglês)

2.3 Recursos e Metodologia Adoptados na Concepção dos Parafraseadores

Os corpora disponibilizados pela Linguateca foram utilizados para a pesquisa de termos e para a análise introspectiva e ilustração de exemplos comprovativos da existência de tais construções. Para além disso, de forma a processar as construções com verbos suporte, o dicionário foi melhorado com propriedades adicionais. A acrescentar à informação mais comum de categoria gramatical e de paradigma flexional, cada entrada do dicionário inclui a descrição dos atributos sintáctico-semânticos (*SynSem*), bem como as propriedades distribucionais e transformacionais para as expressões com um comportamento sintáctico

[1] MTLab é a abreviatura do inglês ***M**achine **T**ranslation **Lab**oratory* (Laboratório de Tradução Automática), um ambiente de desenvolvimento de tradução automática em fase inicial.

mais variável. As entradas apresentam propriedades como: argumentos predicativos, verbos suporte, verbos aspectuais, verbos copulativos, variantes estilísticas dos verbos suporte elementares, informação acerca dos determinantes e preposições que ocorrem com os nomes predicativos em expressões "menos variáveis" e propriedades derivacionais. A derivação é muito importante porque tem implicações não só ao nível lexical, mas também ao nível sintáctico. Muitas vezes, os sufixos derivacionais aplicam-se a palavras de uma categoria sintáctica e transformam-nas em palavras de uma categoria sintáctica diferente, mantendo a sua integridade semântica. Por exemplo, o afixo *-ção* permite transformar o verbo *adaptar* no nome *adaptação* e o afixo *-mente* permite transformar o adjectivo *rápido* no advérbio *rapidamente*. Estas transformações são extremamente importantes para as construções com verbos suporte porque permitem estabelecer gramáticas de equivalência que efectuam o mapeamento entre (i) construções com verbos suporte como *fazer uma adaptação (de)* e o verbo lexical *adaptar*, onde o nome predicativo *adaptação* mantém uma relação semântica e morfossintáctica com o verbo *adaptar* ou (ii) construções com verbos suporte como *ter um final rápido* e a expressão verbal *terminar rapidamente*, onde o nome predicativo autónomo *final* mantém uma relação semântica com o verbo *terminar*, e o advérbio *rapidamente* mantém uma relação semântica e morfossintáctica com o adjectivo *rápido*. Assim sendo, as entradas do dicionário do Port4NooJ contém a identificação dos paradigmas derivacionais para as nominalizações (anotação *NDRV*) e uma ligação ao(s) verbo(s) suporte(s) do nome derivado (anotação *VSUP*), como ilustra a Fig. 4 abaixo. As nominalizações são acompanhadas pelas propriedades correspondentes ao paradigma flexional. Quaisquer outras restrições lexicais, tais como preposições, determinantes, ou argumentos obrigatórios, etc., são igualmente acrescentados. Os nomes predicativos autónomos (não-nominalizações), tais como *favor* são lexicalizados e classificados com a anotação *Npred* e têm associados a eles verbos suporte e outras restrições lexicais, tais como uma preposição (*NPrep*), ou um verbo lexical (*VRB*) com as mesmas características semânticas. Os adjectivos predicativos estão também classificados e foi estabelecida a ligação entre eles e os verbos correspondentes (*ADRV*), tais como entre o verbo *adoçar* e o adjectivo *doce*. Foi também iniciada a atribuição de verbos copulativos (*VCOP*) correspondentes a estes adjectivos. As variantes estilísticas das construções com verbos suportes elementares estão anotadas como *VSTYLE*. As variantes aspectuais estão anotadas como *VASP*. Foi iniciada a adição de argumentos sintácticos e semânticos de um predicado às entradas do dicionário. Por exemplo, na entrada lexical para o verbo *transplantar*, a propriedade *SUBJ=AG* significa que o verbo selecciona um agente como seu argumento semântico na posição sintáctica de sujeito. *SUBJ=PAT* significa que o verbo selecciona um paciente como seu argumento semântico na posição sintáctica de sujeito. O argumento sintáctico *DO=ORG* significa que o predicado selecciona um objecto directo que é um órgão humano (subclasse de parte do corpo). *IO=PAT* significa que o predicado selecciona um objecto indirecto que é um paciente. *NPrep=de* significa que a construção com verbo suporte (verbo suporte mais nome predicativo) selecciona a preposição *de* (*fazer um transplante de*). Os nomes (entidades mencionadas) são classificados semanticamente. Por exemplo, o nome *médico* está classificado como um ser animado que denota uma profissão ou outra designação humana (*AN+des*), pertencente ao domínio médico (*Med*).

adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset132+EN=adapt+*VSUP=fazer*+*DRV*=NDRV00:CANÇÃO +*NPrep*=de
favor,N+FLX=MAR+Npred+AB+state+EN=favor+*VSUP*=fazer+*NPrep*=a+*VRB*=ajudar
rápido,A+FLX=RÁPIDO+PV+eagerType+EN=quick+*DRV*=AVDRV06:RAPIDAMENTE
adoçar,V+FLX=COMEÇAR+Aux=1+OBJTRundif75+Subset604+EN=sweeten+*DRV*=ADRV11:VERDE+*VCOP*=tornar
transplantar,V+FLX=FALAR+Aux=1+RECTR26+Subset=504+BioMed+EN=transplant+SUBJ=AG+VSUP=fazer+DRV=NDRV79:ANO+NPrep=de+DO=BP+IO=PAT+*VSTYLE=sofrer*+*VSTYLE=realizar*+*VSTYLE=efectuar*+*VASP=iniciar*+*VASP=prosseguir*+*VASP=concluir*
médico,N+FLX=ANO+*AN*+*des*+*Med*+EN=doctor
médico,N+FLX=ANO+*AN*+*des*+*Med*+EN=physician

Fig. 4: Amostra do dicionário

As construções com verbos suporte semi-cristalizadas e idiomáticas, onde o verbo suporte é a única palavra que varia em toda a expressão, são lexicalizadas no dicionário de expressões multipalavra e mantidas numa base de dados fraseológica. Por exemplo, em *dar pontadas de dor* ou *pôr cobro a*, na Fig. 5, os verbos suporte *dar* e *pôr* são marcados com uma propriedade correspondente ao paradigma flexional e as restantes palavras na expressão permanecem invariáveis. À medida que os dicionários são melhorados no que respeita à semântica e sintaxe de palavras simples, tenciona-se alargar e redefinir o papel dos dicionários electrónicos de modo a incluir entradas de expressões multipalavra, incluindo construções com verbos suporte e as suas paráfrases.

```
dar parte de fraco,V+SVC+FLX=PHRDAR+EN=become weak+VRB=fraquejar
dar cabo dos nervos,V+SVC+FLX=PHRDAR+EN=enervate+VRB=enervar
dar pontadas de dor,V+SVC+FLX=PHRDAR+EN=hurt+VRB=doer
bater as botas,V+SVC+FLX=PHRBATER+EN=die+VRB=morrer
bater na mesma tecla,V+SVC+FLX=PHRBATER+EN=insist+VRB=insistir
abrir o coração,V+SVC+FLX=PHRABRIR+EN=talk+VRB=desabafar
pôr cobro a,V+SVC+FLX=PHRPOR+EN=end+VRB=terminar
dar lugar a,V+SVC+FLX=PHRDAR+EN=lead to+EN=result in+VRB=conduzir a+VRB=resultar em
dar cabo de,V+SVC+FLX=PHRDAR+EN=destroy+VRB=destruir
pôr um ponto final em,V+SVC+FLX=PHRPOR+EN=end+VRB=acabar com
```

Fig. 5: Amostra da base de dados fraseológica e parafrástica com expressões idiomáticas

O método de reconhecimento e parafraseamento utilizado neste trabalho consiste na ligação sistemática entre palavras relacionadas semântica e morfossintacticamente no dicionário electrónico através do estabelecimento de propriedades derivacionais e distribucionais. De forma a obter as paráfrases monolingues das construções com verbos suporte utilizando o NooJ, combinaram-se as propriedades formalizadas nos dicionários com as gramáticas locais. Uma das novidades deste trabalho em relação ao que já existia, é precisamente a aplicação das gramáticas locais para o reconhecimento e geração de paráfrases de construções com verbos suporte e para a tradução. De modo a estabelecer relações de equivalência morfossintáctica entre predicados nominais e verbais, utilizam-se as propriedades dos dicionários. Uma vez que todos os nomes predicativos estão classificados no dicionário como [*Npred*], esta informação lexical pode ser usada numa gramática local para a identificação do predicado numa construção com verbo suporte e aplicar esta gramática a corpora. A Fig. 6 representa uma gramática local simples usada para reconhecer e gerar construções com verbos suporte e transformá-las nas suas paráfrases verbais.

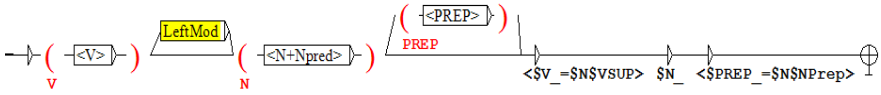


Fig. 6: Gramática para o reconhecimento e parafraseamento de construções com verbos suporte

Esta gramática reconhece verbos suporte seguidos de um modificador (determinante, adjectivo, advérbio ou outros quantificadores), de um nome predicativo e opcionalmente de uma preposição. Os elementos entre parênteses ( ) são guardados em variáveis V, N ou PREP. Se uma entrada de dicionário contém uma restrição lexical, tal como NPrep=a na expressão [*dar um grande abraço a*], a construção com verbo suporte será reconhecida pela gramática e mapeada ao verbo *abraçar*, o lema do nome especificado na variável $N_. Os elementos a negrito <$V_=$N$VSUP>, e $PREP_=$N$NPrep> representam restrições lexicais que são apresentadas na saída, tais como a especificação do verbo suporte ou da preposição que pertence a uma construção com verbo suporte específica. O nome predicativo é identificado, mapeado com o seu derivante e exibido como um verbo. Os outros elementos da expressão são eliminados.

2.4 Avaliação Quantitativa: Primeiros Resultados

Para a avaliação do ReWriter, foram seleccionadas a partir do Compara todas as frases onde a forma infinitiva dos verbos *fazer*, *dar*, *pôr*, *tomar* e *ter* ocorria com um nome ou com um determinante e um nome. Em primeiro lugar, foram classificadas manualmente estas combinações para ver se elas correspondiam a construções com verbos suporte ou não. Confirmou-se que 89% das ocorrências de *dar*, 88% de *tomar*, 77% de *pôr*, 47% de *fazer* e 20% de *ter* são verbos suporte. Isto significa que na sua globalidade, em 64.2% das vezes estes verbos são verbos suporte, o que corresponde a quase 2/3 das ocorrências. A seguir a esta contagem, foi seleccionado um sub-corpus de 500 frases obtidas de modo aleatório (100 frases para cada um dos cinco verbos seleccionados), contendo apenas construções com verbos suporte. As construções foram anotadas manualmente e os resultados comparados com os resultados obtidos automaticamente. Elaboraram-se regras de reconhecimento mais restritas para que o parafraseamento fosse mais preciso. Actualmente, são reconhecidas 62.6% de construções com verbos suporte com valores elevados em termos de precisão. Os resultados do reconhecimento e parafraseamento (precisão e cobertura) do ReWriter estão ilustrados na Fig. 7.

	Reconhecimento Precisão	Reconhecimento Cobertura	Parafraseamento Precisão
Pôr	73/73 - 100%	73/100 - 73%	72/73 - 98.6%
Tomar	75/75 - 100%	75/100 - 75%	68/73 - 93.1%
Ter	65/65 - 100%	65/100 - 65%	59/65 - 90.7%
Dar	57/60 - 95%	57/100 - 57%	46/51 - 90.1%
Fazer	43/45 - 95.5%	43/100 - 43%	40/45 - 88.8%
Média	62.6/63.6 - **98.4%**	62.6/100 - **62.6%**	57/61 - **93.4%**

Fig. 7. Avaliação do reconhecimento e parafraseamento de construções com verbos suporte

3 Considerações Finais

Os parafraseadores ReWriter e ParaMT e os recursos linguísticos do Port4NooJ, que estão na base destas ferramentas, podem ser integrados facilmente noutros recursos da Linguateca e colocados ao serviço da comunidade. Os recursos do Port4NooJ já estão a ser utilizados no Corpógrafo, mas a sua versatilidade e detalhe linguístico, nomeadamente a informação sintáctica e semântica são apropriados para obtenção de concordâncias mais sofisticadas e extracção de termos, expressões multipalavra e fraseologia. Prevê-se a criação de um maior número de gramáticas de desambiguação para análise sintáctico-semântica e o desenvolvimento de dicionários mais completos e mais ricos em informação linguística. Há também o objectivo de criar interfaces de acesso público que permita um uso interactivo do ReWriter e do ParaMT. O passo seguinte será a utilização dos novos recursos para testar e melhorar estas aplicações, que servirão posteriormente para uma anotação mais completa rigorosa dos corpora anotados, como por exemplo, do AC/DC. E finalmente, o alargamento dos recursos, de modo a desenvolver o sistema de tradução automática já iniciado. A falta de projectos de tradução automática envolvendo o português, deixa a nossa língua desfasada da realidade da tradução automática e é necessário colmatar esta deficiência através de iniciativas como as que já foram propostas no âmbito da Linguateca.

A política de disponibilização e partilha de recursos praticada pela Linguateca, a colaboração e junção de esforços começa agora a gerar os seus primeiros frutos. É importante salvaguardar os recursos até ao momento produzidos, mantendo-os em sistemas de fácil acesso, como em código aberto. Como as peças de um puzzle que se vão unindo para formar um todo, é necessário juntá-los para que se criem a partir deles recursos cada vez maiores, mais completos e mais enriquecidos linguisticamente. Estão criadas as infra-estruturas e reunidas as competências e condições necessárias para a criação de colaborações que possam ter objectivos concretos em relação aos actuais desafios tecnológicos de um mundo cada vez mais virado para a globalização da informação. É importante criar iniciativas semelhantes à da Linguateca, e até mesmo, há necessidade de criar um organismo ou uma sociedade internacional de análise e processamento de língua portuguesa, com actividades centradas em áreas específicas, mas sempre com uma visão global da língua. A especialização de recursos humanos em várias áreas do processamento do português, nomeadamente em entidades mencionadas, entidades geográficas, ontologias, extracção e recuperação de informação, tradução automática, entre outras, são uma mais-valia que deve ser aproveitada pela sociedade em geral, tanto para o desenvolvimento de ferramentas de utilidade pública como privada. Como legado da Linguateca, podemos contar com uma nova etapa para o futuro do processamento da língua portuguesa, com novos desafios e inúmeras oportunidades!

Referências

[1] Susana Afonso, Eckhard Bick, Renato Haber & Diana Santos. "Floresta sintá(c)tica: um treebank para o português". In Anabela Gonçalves & Clara Nunes Correia (eds.), *Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística (APL 2001)* (Lisboa, 2-4 de Outubro de 2001), Lisboa: APL, pp. 533-545.

[2] Susana Afonso, Eckhard Bick, Renato Haber & Diana Santos. "Floresta sintá(c)tica: a treebank for Portuguese". In Manuel González Rodrigues & Carmen Paz Suarez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation (LREC 2002)* (Las Palmas de Gran Canaria, Espanha, 29-31 de Maio de 2002), Paris: ELRA, pp. 1698-1703.

[3] Diana Santos & Anabela Barreiro. "On the problems of creating a consensual golden standard of inflected forms in". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation (LREC'2004)* (Lisboa, Portugal, 26-28 de Maio de 2004), pp. 483-486.

[4] Anabela Barreiro & Susana Afonso. "Construção da lista dourada para as primeiras Morfolimpíadas do português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 107-118.

[5] Luís Sarmento, Anabela Barreiro, Belinda Maia & Diana Santos. "Avaliação de Tradução Automática: alguns conceitos e reflexões". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 181-190. 295

[6] Belinda Maia & Anabela Barreiro. "Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 205-216.

[7] Luís Sarmento. "Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática". In D. Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal: IST Press, 2007, pp. 193-203.

[8] Santos, D., B. Maia, L. Sarmento. "Gathering empirical data to evaluate MT from English to Portuguese". In Lambros Kranias, Nicoletta Calzolari, Gregor Thurmair, Yorick Wilks, Eduard Hovy, Guðrún Magnúsdóttir, Anna Samiotou & Khalid Choukri (eds.), *Proceedings of LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora* (Lisboa, Portugal, 25 de Maio de 2004), pp. 14-17.

[9] Ana Frankenberg-Garcia & Diana Santos. "Introducing COMPARA, the Portuguese-English parallel translation corpus". In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in Translation Education*. Manchester: St. Jerome Publishing, 2003, pp. 71-87. http://www.linguateca.pt/COMPARA/

[10] Anabela Barreiro. "Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation". In Xavier Blanco and Max Silberztein (eds). *Proceedings of the 2007 International NooJ Conference*. Univ. Autonoma de Barcelona, June 7-9, 2007. Cambridge Scholars Publishing, 2008 (forthcoming).

[11] http://www.nooj4nlp.net/

[12] http://www.linguateca.pt/Repositorio/Port4Nooj/

[13] Max Silberztein. "NooJ: A Cooperative, Object-Oriented Architecture for NLP". In *INTEX pour la Linguistique et le traitement automatique des langues*. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté, 2004. http://www.nooj4nlp.net/

[14] Luís Sarmento, Belinda Maia & Diana Santos. "The Corpógrafo - a Web-based environment for corpora research". In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of LREC2004, the 4th International Conference on Language Resources and Evaluation (LREC'2004)* (Lisboa, Portugal, 26-28 May 2004), pp. 449-452. http://www.linguateca.pt/Corpografo/

[15] Belinda Maia & Sérgio Matos. "Corpografo V4 - Tools for Researchers and Teachers using Comparable Corpora". In Pierre Zweigenbaum, Éric Gaussier & Pascale Fung (eds.), *LREC 2008 Workshop on Comparable Corpora (LREC 2008)* (Marrakech, 31 May, 2008), European Language Resources Association (ELRA), pp. 79-82.

[16] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, Ralph Grishman. "Annotating noun argument structure for NomBank". *Proceedings of LREC 2004*. Lisbon, Portugal.

[17] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekeley, Veronique Zielinska, Brian Young. "The Cross-Breeding of Dictionaries". *Proceedings of LREC-2004*, Lisbon, Portugal.

[18] Maurice Gross. "Les bases empiriques de la notion de prédicat sémantique". In A. Guillet and C. Leclère (eds). *Formes Syntaxiques et Prédicat Sémantiques*, *Langages*, 63: 7-52. Larousse, Paris. 1981.

[19] Maurice Gross. *Méthodes en syntaxe*. Hermann. 1975.

[20] Zellig Harris. *Mathematical Structures of Language*, New York: Wiley, 230p. 1968.

[21] Zellig Harris. "Co-occurrence and transformation in linguistic structure". *Language*, 33, 293-340. 1957.

[22] Elisabete Ranchood. *Sintaxe dos predicados nominais com Estar*, 477p., Lisboa: INIC, 1990.

[23] Jorge Baptista, *Sintaxe dos predicados nominais construídos com* Ser de, Universidade do Algarve, 2001.

[24] Lucília Chacoto. *O Verbo Fazer em Construções Nominais Predicativas.* Universidade do Algarve, 2005.

[25] Anabela Barreiro. "ParaMT: a Paraphraser for Machine Translation". In António Teixeira, Vera Strube de Lima, Luís Caldas de Oliveira, Paulo Quaresma (eds). *Proceedings of the International Conference on Computational Processing of Portuguese*, PROPOR 2008, LNAI 5190. Universidade de Aveiro, September, 8-10, 2008. Springer, LNCS/LNAI.