Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

O processo de avaliação do sistema ConPor

Lucia Specia
Lucia Helena Machado Rino

NILC-TR-03-05

Fevereiro, 2003

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

Resumo

Este relatório descreve o processo de avaliação do sistema **ConPor**, cuja função é mapear estruturas sintáticas em estruturas conceituais UNL. São apresentados dois testes específicos realizados para avaliar o desempenho do sistema, em termos da qualidade das regras de mapeamento e da expressividade das estruturas conceituais geradas, com relação ao significado das correspondentes sentenças originais.

Este trabalho conta com o apoio
financeiro da CAPES

Índice

Figuras

Tabelas

1 Introdução

Nos últimos anos, a avaliação da qualidade de sistemas de Processamento de Línguas Naturais (PLN) tem assumido um papel de grande importância, uma vez que esses sistemas deixaram de ser apenas protótipos de modelos para se tornarem aplicações reais, voltadas para usuários comuns. Avaliações substanciais indicam o progresso dos sistemas e das próprias pesquisas na área de PLN, além de permitirem comparar diferentes metodologias e sistemas para uma determinada tarefa. Como as diversas aplicações do PLN focalizam diferentes tarefas, devem ser submetidas a diferentes procedimentos de avaliação. Para algumas aplicações, existem procedimentos padronizados, que definem, entre outras coisas, o conjunto de dados a ser avaliado, a configuração da avaliação (formato dos dados, gênero e/ou domínio textual, etc.), as características que devem ser avaliadas, o método de avaliação, o conjunto de resultados esperados e os valores de referência para classificar os resultados obtidos. Os analisadores sintáticos são um exemplo de aplicação para a qual existem padronizações dessa natureza, como o conjunto de dados a ser avaliado e o de resultados esperados, ambos especificados no *corpus* Penn Treebank (Marcus et al., 1994). Esse tipo de padronização permite a avaliação constrastiva dos resultados de diferentes sistemas, além da avaliação progressiva do próprio sistema.

No caso das aplicações para as quais não há procedimentos padronizados, como aquelas que envolvem a análise semântica, foco deste relatório, normalmente são realizadas avaliações autônomas, sem comparação com os resultados de outros sistemas. Neste caso, a avaliação consiste da medida da própria qualidade da aplicação, contemplando características muito peculiares a ela.

O objetivo deste relatório é descrever o processo empírico empregado para a avaliação autônoma do desempenho do sistema ConPor, o gerador de estruturas conceituais para sentenças em português, já descrito anteriormente (Specia & Rino, 2002a; 2003). Nesse processo são realizados dois testes, visando avaliar: (a) a qualidade das regras de geração de estruturas conceituais; e (b) o poder de expressividade dessas estruturas, com relação ao significado das sentenças originais correspondentes. Para contextualizar tal processo, na Seção 2 apresentamos as principais características do sistema ConPor. Na Seção 3 ilustramos algumas possibilidades de configuração da avaliação de sistemas de PLN. Na Seção 4 abordamos algumas iniciativas de avaliação de sistemas de interpretação semântica, em particular, para descrever, na Seção 5, a avaliação do ConPor. Na Seção 6 apresentamos algumas considerações finais.

2 O sistema ConPor

O **ConPor** (**Con**ceitualização do **Por**tuguês) é um sistema de interpretação sentencial do português cuja tarefa consiste em gerar estruturas conceituais UNL (*Universal Networking Language*) (UNL, 2001) a partir de estruturas sintáticas produzidas pelo *parser* Curupira, em desenvolvimento no NILC[1]. Essa tarefa de mapeamento de estruturas sintáticas em estruturas conceituais (**mapeamento sintático-conceitual**, doravante) é realizada pelo módulo **Gerador Conceitual** (Figura 1), a partir do conhecimento lingüístico armazenado no Repositório Conceitual e no Léxico Enriquecido.

[1] Núcleo Interinstitucional de Lingüística Computacional (www.nilc.icmc.sc.usp.br).

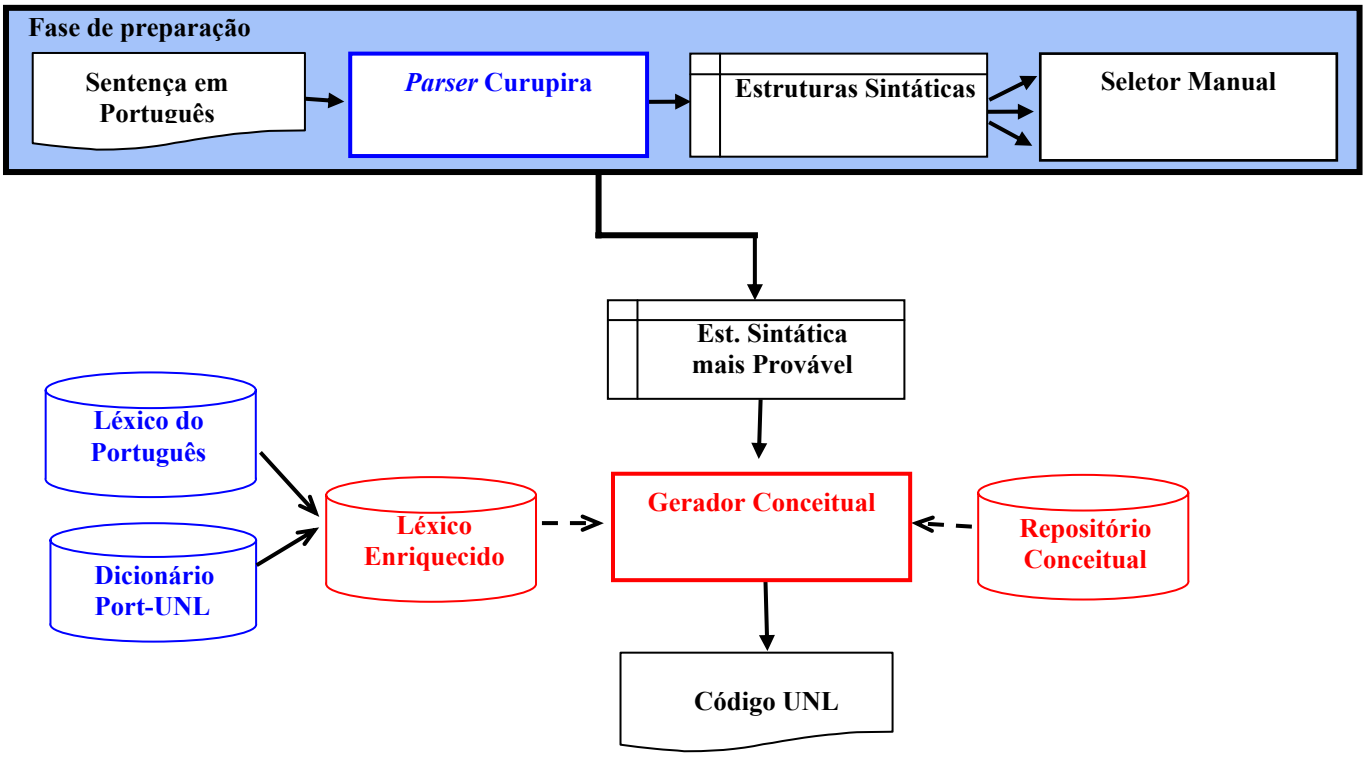

Figura 1 – Arquitetura do ConPor

A seguir, sintetizamos as características do ConPor relevantes ao processo de avaliação.

2.1 Dados de entrada

A entrada efetiva para o processo de mapeamento consiste da estrutura sintática de uma sentença do português, escolhida manualmente dentre todas aquelas geradas pelo *parser*. Como delimitação de projeto, o sistema contempla somente um subgrupo de construções gramaticais no domínio de textos de horóscopo. Seus recursos lingüísticos e processos foram construídos com base nas características de um conjunto de sentenças extraídas da coluna de horóscopo diário do jornal *on-line* Folha de São Paulo[2], constituindo o chamado *corpus* base. Essas sentenças obedecem, basicamente, a dois critérios: (a) são sentenças simples (com um único verbo ou locuções verbais) ou, no máximo, sentenças subordinadas reduzidas; e (b) não apresentam ambigüidade semântica, ou seja, se uma palavra é utilizada com um dado sentido em uma sentença, não pode ser utilizada com sentido diferente em outra sentença, a não ser que as palavras sejam de categorias gramaticais diferentes.

O primeiro critério visa delimitar o escopo inicial do trabalho para casos mais simples, de acordo com uma metodologia de construção incremental. O segundo critério visa evitar problemas de escolha lexical, uma vez que o objetivo principal do sistema é o mapeamento de relações sintáticas (da estrutura sintática) em relações semânticas (da estrutura conceitual), e

[2] http://www1.folha.uol.com.br/folha/urania/

não a escolha lexical, isto é, a seleção dos conceitos que correspondem às palavras da sentença.

2.2 Recursos

O Léxico Enriquecido foi criado a partir de informações de outros dois recursos, o Léxico do Português (Nunes et al., 1996) e o Dicionário Português UNL (Dias-da-Silva et al., 1998). Esse léxico é personalizado para o gênero e domínio considerados, ou seja, sentenças de horóscopo diário. Para evitar a ambigüidade de sentido na geração conceitual, representamos nesse léxico somente uma acepção para cada palavra, ou seja, há somente uma entrada para cada palavra da mesma categoria gramatical e, com isso, cada palavra das sentenças do *corpus* é mapeada em um único conceito UNL, como já citamos. Detalhes sobre sua construção podem ser consultados em Specia & Rino (2002b).

O Repositório Conceitual fornece as regras para o mapeamento sintático-conceitual, especificadas no formato de *templates*, divididas em duas categorias: (a) regras de projeção, para mapear os constituintes sintáticos em constituintes conceituais, associados aos seus papéis semânticos, seguindo uma abordagem gerativo-transformacional; e (b) heurísticas de relacionamento, para combinar cada par de constituintes conceituais resultantes da aplicação das regras de projeção em uma relação binária UNL, de acordo com seus papéis semânticos. As heurísticas são definidas de acordo com um mecanismo suficientemente claro e conciso para evitar relacionamentos incorretos ou ambíguos.

2.3 Geração conceitual

O processo de geração conceitual, propriamente dito, é realizado pelo mecanismo de inferência do módulo Gerador Conceitual. Esse processo ocorre a partir da aplicação de um determinado número de regras do Repositório Conceitual sobre a estrutura sintática de entrada para o mapeamento das suas funções sintáticas em relações semânticas UNL e das suas palavras em conceitos UNL.

2.4 Dados de saída

A estrutura conceitual da linguagem UNL é uma representação lógica, baseada em relações semânticas binárias entre conceitos independentes de língua. Para tanto, a UNL dispõe de três componentes: rótulos de relações semânticas (RLs – *Relation Labels*), palavras que indicam conceitos universais (UWs – *Universal Words*) e rótulos de atributos que especificam ou restringem esses conceitos (ALs – *Attribute Labels*). Com base nesses componentes, uma estrutura conceitual UNL é constituída de uma ou mais relações com o formato abaixo, sendo que UW_1 e UW_2 devem ser diferentes e os ALs são opcionais e em número variável, sempre precedidos por '@'.

$$RL(UW_1.@AL, UW_2.@AL)$$

Como um exemplo, considere a sentença (1) e sua representação UNL na Figura 2, que utiliza quatro relações semânticas (agt = agente; plc = lugar; obj = objeto; pos = posse), envolvendo os conceitos *sun* (sol), *sagittarius* (sagitário), *illuminate* (ilumina) e *relationship*

(relacionamentos). Alguns desses conceitos têm atributos associados a eles, como *@pl* em *relationship*, que indica que o conceito é coletivo.

(1) O Sol em Sagitário ilumina seus relacionamentos.

```
agt(illuminate.@entry, sun.@def)
plc(sun.@def, sagittarius)
obj(illuminate.@entry, relationship.@pl)
pos(relationship.@pl, you)
```

Figura 2 – Código UNL para a sentença (1)

3 Configurações para a avaliação de sistemas de PLN

Dependendo dos objetivos pretendidos, os sistemas de PLN podem utilizar diferentes configurações para seus processos de avaliação. Nesta seção, apresentamos as opções mais comuns de configuração, de modo a contextualizar aquela especificada para o ConPor.

Primeiramente, é importante ressaltar que as configurações citadas aqui são voltadas para a avaliação de **sistemas de PLN fundamentais**, isto é, baseados em conhecimento (lingüístico e/ou extralingüístico), e não de **sistemas de PLN empíricos**, isto é, baseados em técnicas estatísticas ou empíricas.

Outra questão importante é que a forma de avaliação que estamos tratando diz respeito ao **desempenho do sistema** em termos da qualidade dos seus resultados. Assim, outras formas de avaliação que meçam, por exemplo, a facilidade de uso, a modularidade, a portabilidade, a manutenibilidade, a eficiência, o custo ou a complexidade necessários para realizar determinada tarefa não são consideradas.

De um modo geral, segundo Sparck-Jones & Galliers (1996), o desempenho dos sistemas de PLN pode ser avaliado sob dois aspectos: com relação ao seu objetivo e com relação ao seu uso. A primeira forma de avaliação é chamada de **avaliação intrínseca** e com ela analisa-se o "sistema em si", ou seja, analisam-se os seus resultados de acordo com o objetivo específico para o qual ele foi desenvolvido. No caso de um sistema de interpretação semântica, avaliam-se as estruturas de representação do significado (representações ou estruturas semânticas ou conceituais) produzidas. A segunda forma de avaliação é chamada de **avaliação extrínseca** e com ela analisa-se o "sistema em uso", ou seja, sua contribuição (como módulo de outro sistema, por exemplo) para os resultados de alguma tarefa específica. Um sistema de interpretação semântica poderia, por exemplo, ser avaliado extrinsecamente no contexto da tradução automática. Vale ressaltar que ambas as formas de avaliação, intrínseca e extrínseca, não são mutuamente exclusivas. Na verdade, podem ser consideradas complementares, uma vez que avaliam diferentes características dos sistemas.

Outra diferenciação nos processos de avaliação, utilizada, por exemplo, por Sparck-Jones & Galliers, (1996) e Palmer & Finin (1990), é a que distingue a **avaliação *blackbox*** da **avaliação *glassbox***. A avaliação *glassbox* é realizada sobre os componentes internos do sistema, analisando seus resultados intermediários. A avaliação *blackbox*, por sua vez, é realizada somente sobre os dados de saída, com base em um conjunto de dados de entrada, ou seja, mede o desempenho do sistema em termos de pares bem definidos de entrada e saída. Avaliações *glassbox* só são possíveis quando existem componentes distinguíveis associados a diferentes estágios de processamento, os quais podem, então, ser avaliados isoladamente. Por exemplo, em um sistema de interpretação semântica, avaliações *glassbox* poderiam incluir o

exame da identificação dos papéis semânticos, dos conceitos, das relações semânticas, etc, enquanto uma avaliação *blackbox* mediria a qualidade das representações conceituais resultantes.

Os dados de entrada para o processo de avaliação do sistema podem ser especificados de duas formas: com base em um conjunto de textos autênticos, de gênero e domínio determinados, denominado ***corpus* de teste**, ou com base em um conjunto de dados construído artificialmente, visando contemplar aspectos específicos a avaliar, denominado **conjunto seletivo de teste** (*test suite*) (Krenn & Samuelsson, 1997).

Os resultados automáticos podem ser avaliados de forma comparativa, com relação a **resultados de referência**, também chamados aqui de resultados ideais ou *gold standards* (Teufel & Moens, 1999), ou com relação aos resultados mínimos esperados, chamados comumente de **medida de base**, ou, simplesmente, ***baseline***. Os resultados de referência normalmente são produzidos manualmente, por especialistas no domínio e/ou na tarefa em foco, ou, simplesmente, por falantes nativos da língua natural (LN) em foco no sistema. A *baseline* pode ser produzida tanto por humanos quanto por outros sistemas correlatos. Os sistemas também podem ser avaliados subjetivamente, por meio de procedimentos de análise isolada, não comparativa, executada por **juízes humanos**. Os juízes humanos são, em geral, leitores fluentes da LN em foco, podendo ou não ser conhecedores do assunto contemplado pelos dados. Escolhas como essas dependem da tarefa de avaliação, do resultado a ser avaliado e também da disponibilidade de dados (ideais ou *baseline*) para comparação.

Além dessas configurações gerais, a execução do procedimento de avaliação depende de três fatores principais: os **critérios** que devem ser avaliados, as **medidas/métricas** para verificar se os critérios são satisfeitos, e os **métodos** utilizados para essa avaliação (Sparck-Jones & Galliers, 1996; Hirschman & Thompson, 1996).

Um critério indica o que se pretende medir com a avaliação, por exemplo, a velocidade de processamento, a qualidade de uma tradução, a habilidade para recuperar documentos relevantes, a abrangência do sistema, etc. Uma medida expressa a propriedade específica do desempenho do sistema que será utilizada para avaliar determinado critério, por exemplo, para o critério "velocidade de processamento", o número de segundos por processo, para o critério "qualidade", a precisão (soluções corretas encontradas pelo sistema em relação a todas as soluções encontradas por ele), a cobertura (soluções corretas encontradas pelo sistema em relação a todas as soluções corretas), a taxa de erro (soluções corretas não encontradas pelo sistema) e a taxa de falha (soluções incorretas encontradas pelo sistema), etc. Um método indica como determinar as medidas, por exemplo, se por *benchmark*, por julgamento por humanos, etc. Por exemplo, na área de recuperação de informação, um critério clássico é a habilidade de um sistema de recuperar documentos relevantes, dada uma descrição do tópico. Uma medida para esse critério é o número de documentos recuperados que são relevantes, ou seja, é a cobertura do sistema. Um método para computá-la consiste em calcular a divisão do número de documentos recuperados relevantes por todos os documentos recuperados.

Com relação à metodologia, a realização dos testes pode ser **manual**, **automática** ou **semi-automática**. Os testes automatizados, quando possíveis, permitem **avaliações objetivas**. Já os testes manuais, feitos pela própria equipe de desenvolvimento do sistema ou por juízes humanos externos, especialistas ou não em determinada tarefa e/ou domínio, constituem **avaliações subjetivas**, dependentes do julgamento ou da experiência dos próprios juízes.

No caso de avaliações subjetivas, a representatividade da avaliação depende do número de juízes envolvidos. No entanto, em função do próprio caráter subjetivo, é preciso considerar a variação analítica dos julgamentos, de modo a garantir que ela não interfira na

qualidade da própria avaliação. Uma forma de fazer isso é medir o grau de variação da resposta dos juízes, de modo a verificar se as suas opiniões não diferem tão profundamente entre si a ponto de invalidar o teste. Para tanto, existem medidas de confiabilidade para testes com juízes humanos, entre elas, o índice **Kappa** (Siegel & Castellan, 1977), que mede a concordância dos juízes em todas as classes de respostas e a **concordância por classe** (Poesio & Vieira, 1998), que mede a concordância dos juízes para cada classe de respostas.

Por fim, conforme mencionamos, os resultados obtidos com os testes na avaliação de um sistema podem ser analisados isoladamente (**avaliação autônoma**) ou em comparação com resultados da avaliação de outros sistemas afins (**avaliação comparativa**) (Palmer & Finin, 1990). Atualmente, são poucas as subáreas do PLN para as quais são realizadas iniciativas de avaliações conjuntas, dentre as quais destacamos aquelas delineadas pelas conferências MUC (*Message Understanding Conference*) (por exemplo, MUC-7, 1998), para aplicações de compreensão de mensagens, as conferências TREC (*Text Retrieval Conference*)[3], para aplicações de recuperação de informações, as conferências DUC (*Document Understanding Conference*)[4] para sumarização automática, e os Workshops de Avaliação em Tradução Automática (EAGLES e MT Summit, por exemplo)[5].

Várias das características de configuração apresentadas nesta seção são incorporadas à avaliação do ConPor. Antes de descrevê-la, no entanto, na próxima seção apresentamos algumas outras iniciativas específicas para a avaliação de sistemas de interpretação semântica.

4 A avaliação de sistemas de interpretação semântica

4.1 Avaliações extrínsecas comparativas

Os pesquisadores da área de interpretação semântica, em função da crescente importância atribuída às avaliações comparativas entre sistemas, vêm buscando o desenvolvimento de procedimentos padronizados de avaliação, por meio de esforços como as conferências MUC, realizadas pela DARPA (*Defense Advanced Research Projects Agency*). Nessas conferências, os sistemas são avaliados segundo seu desempenho na realização de algumas tarefas específicas, como a extração de informações, a resolução de co-referências e o reconhecimento de nomes de entidades. Portanto, para que quaisquer aplicações possam ser avaliadas no contexto da MUC, é preciso que elas sejam desenvolvidas ou adaptadas para contemplar tais tarefas.

Apesar de demonstrarem a aplicabilidade dos sistemas para determinadas tarefas, avaliações extrínsecas como as realizadas pela DARPA não permitem analisar características específicas dos sistemas de interpretação semântica, como a qualidade das estruturas semânticas, pois o sistema é visto como um todo, produtor de resultados que não serão julgados diretamente pelo desempenho do sistema, mas por sua validade, aplicabilidade ou pertinência na realização de alguma tarefa.

[3] http://trec.nist.gov.
[4] http://www-nlpir.nist.gov/projects/duc/.
[5] http://www.cst.dk/projects/eagles2/othersbody.html e http://www.eamt.org/summitVIII, respectivamente.

4.2 Avaliações intrínsecas comparativas

A avaliação comparativa das estruturas semânticas é um problema bastante complexo, em função da própria complexidade da tarefa de interpretação semântica, pois uma mesma expressão pode ter múltiplas interpretações e essas interpretações podem ser representadas de múltiplas formas. De fato, os sistemas de interpretação semântica podem seguir diferentes filosofias de projeto e, com isso, utilizar diferentes teorias semânticas, que privilegiam a representação de diferentes aspectos da LN e do significado, em diferentes formas de representação, indicando diversas formas de processamento. Dessa forma, ainda que os sistemas tenham o mesmo objetivo e que visem tratar do mesmo gênero e domínio textual, torna-se difícil estabelecer padrões para os procedimentos de avaliação, assim como para conjuntos de dados de entrada e seus respectivos resultados de referência.

Apesar dessas dificuldades, alguns esforços vêm sendo empregados, já há algum tempo, na tentativa de desenvolver procedimentos de avaliação intrínseca comparativa. Já em 1990, o Workshop de Avaliação de Sistemas de PLN (Palmer & Finin, 1990), por exemplo, teve como uma das premissas justamente a investigação de métodos de avaliação genéricos para interpretadores semânticos. Todavia, por conta da falta de consenso sobre como a qualidade das representações semânticas deveria ser avaliada, as conclusões do Workshop foram de que a única forma de avaliação comparativa viável seria aquela relacionada a alguma tarefa específica.

Outras iniciativas na tentativa de padronização da avaliação de estruturas semânticas incluem a do projeto SEMEVAL (Moore, 1994), que propõe um *corpus* anotado com etiquetas semânticas, como uma extensão da abordagem similar para representações sintáticas (PARSEVAL) (Black et al., 1991). Sobre essas iniciativas, no entanto, incidem todos os problemas de falta de consenso citados, principalmente a falta de concordância sobre o conjunto de categorias semânticas que devem ser contempladas na avaliação e sobre os sentidos das palavras que devem ser considerados no léxico do sistema.

Após esses trabalhos, relativamente antigos, pouco se tem avançado no campo da avaliação intrínseca comparativa, justamente em função da complexidade para se identificar as características mais relevantes para tal tarefa de avaliação, que sejam comuns a todos os sistemas a serem avaliados. Dadas essas dificuldades, as avaliações intrínsecas, focalizando especificamente as estruturas de representação do significado, geralmente são feitas de maneira autônoma, isto é, considerando um único sistema por vez. Neste caso, privilegiam-se critérios de avaliação e medidas e métricas correspondentes personalizados para a natureza da representação semântica e os objetivos em foco.

4.3 Avaliações intrínsecas autônomas

Dentre as poucas abordagens de avaliação intrínseca encontradas, algumas são relevantes para este trabalho, pois consideram configurações de teste de desempenho interessantes ao ConPor. Este é o caso das metodologias propostas por Bean et al. (1998), Romacker & Hahn (2000) e Rosé (2000). As duas primeiras focalizam a qualidade das representações semânticas, avaliando sua cobertura e precisão. Rosé, por sua vez, focaliza a qualidade dos resultados decodificados para a LN, quando comparados com as sentenças originais correspondentes, avaliando a fluência desses resultados e o quão bem eles comunicam a idéia central das sentenças originais. Quanto ao escopo das avaliações, Bean et al. avaliam apenas as representações semânticas de relações espaciais no campo de anatomia.

Romacker & Hahn avaliam as representações semânticas em diversos domínios, contemplando, contudo, somente dois aspectos da língua inglesa: os casos genitivos e os auxiliares modais. Rosé avalia diversos tipos de representações semânticas do domínio de agendamento de reuniões.

Como veremos na próxima seção, a avaliação do ConPor contempla ambos os critérios citados.

5 A avaliação do ConPor

Para avaliar o desempenho do ConPor, foram especificados dois testes, o primeiro focalizando a qualidade das regras de mapeamento sintático-conceitual e o segundo focalizando o poder de expressividade das estruturas conceituais geradas, com relação ao significado das sentenças originais correspondentes, ou seja, a qualidade das potenciais sentenças em português, resultantes das estruturas semânticas produzidas automaticamente pelo ConPor. Esses dois testes são descritos a seguir.

5.1 Configuração da avaliação

A seguinte configuração geral de testes é considerada:

(1) Avaliação intrínseca: optamos apenas pela análise da qualidade e representatividade das estruturas conceituais, independentemente da sua aplicação, porque nosso sistema (a) ainda não é um sistema completo e, portanto, não serve a nenhuma outra tarefa, diretamente e (b) nosso interesse é, sobretudo, verificar se o mapeamento sintático-conceitual está adequadamente contemplado e se as estruturas conceituais geradas são consideradas satisfatórias, quando comparadas aos dados de entrada.

(2) Avaliação *blackbox*: definimos como objeto de avaliação somente o resultado final do mapeamento, ou seja, as estruturas conceituais ou códigos UNL gerados.

(3) Avaliação autônoma: os resultados obtidos com os testes não são comparados com resultados de outros sistemas, uma vez que não há procedimentos padronizados para tal avaliação, em função de todas as dificuldades já citadas na Seção 4.2.

(4) Definição do *corpus* de teste: definimos os dados de teste do sistema de acordo com um *corpus* de teste, no mesmo domínio do *corpus* base: o de textos de horóscopo diário. Tais textos foram coletados a partir de 10 fontes diferentes, incluindo portais da Internet, revistas e jornais *on-line*, em um mesmo dia, resultando em um total de 360 sentenças. Para compor o *corpus* de teste, entretanto, foram selecionadas somente 80 sentenças (correspondendo a 60 estruturas sintáticas diferentes), uma vez que várias das sentenças originais não seguiam o padrão sintático previsto pelo ConPor e foram, portanto, excluídas.

Como o ConPor não trata a ambigüidade de sentido (conforme mencionado na Seção 2), as palavras ambíguas no *corpus* de teste foram substituídas por equivalentes (sinônimos) não ambíguas. A reescrita do *corpus* não o invalida, do ponto de vista semântico, embora estejamos alterando sua autenticidade. Consideramos que tal alteração lexical não implica alteração conceitual significativa porque (a) o leitor humano também desambigüisa as palavras e, por simplificação, executamos esse processo antes do mapeamento sintático-conceitual; (b) a desambiguação humana é válida e correta, pelo contexto, pois as palavras substituídas têm correspondentes licenciados no léxico da própria língua portuguesa; (c) muito embora isso restrinja o escopo de atuação do ConPor, a limitação do sistema aos casos

não ambíguos ainda permite verificar sua capacidade de gerar estruturas conceituais pertinentes, em UNL; e (d) considerando o gênero do texto, de linguagem simples, domínio restrito e público alvo-genérico, as ocorrências de ambigüidade não são significativas, ou seja, as palavras normalmente são utilizadas sempre em uma única acepção e, com isso, a quantidade de ambigüidades é muito pequena.

(5) *Corpus* de referência: as sentenças do *corpus* de teste foram submetidas à codificação manual, por um especialista em UNL, que não teve acesso à metodologia de codificação do sistema, ou seja, às suas regras de mapeamento. Desse processo resultou o *corpus* de referência, de códigos UNL considerados ideais, isto é, com representações conceituais adequadas. Esse *corpus* foi codificado manualmente porque não dispúnhamos de um codificador Português-UNL automático.

Além dessa configuração geral, as demais decisões do processo de avaliação do ConPor são dependentes dos objetivos de cada um dos testes realizados, e serão apresentadas na próxima seção.

5.2 O processo de avaliação

5.2.1 Teste 1 – Avaliação das regras de mapeamento

O objetivo desse primeiro teste é avaliar o desempenho das regras de mapeamento sintático-conceitual, ou seja, verificar se tais regras geram estruturas conceituais adequadas, quando comparadas com as estruturas do *corpus* de referência. Para tanto, o teste foi configurado como segue:

Critério: desempenho das regras de mapeamento.

Medidas adotadas: cobertura, precisão e *f-measure* para cada estrutura conceitual, aqui denominadas **medidas individuais**; valores médios das mesmas medidas, agora considerando todas as estruturas conceituais (denominadas **médias**).

Método: comparação manual dos códigos do *corpus* de referência com os códigos gerados pelo ConPor, para as 80 sentenças do *corpus* de teste. Essa comparação foi realizada por um único juiz humano, que verificou se o conjunto de relações binárias UNL de cada estrutura conceitual gerada automaticamente coincidia com o conjunto de relações binárias da estrutura conceitual do *corpus* de referência, atribuindo, para cada par de relações binárias, um dos seguintes escores: **sucesso** (para relações idênticas), **falha** (para relações excedentes) e **omissão** (para relações não geradas). Somente pares de relações idênticas, ou seja, com o mesmo rótulo de relação (RL), envolvendo os mesmos conceitos (UWs), são consideradas coincidentes[6]. Optamos por esse tipo de avaliação estrita, como sugerem Romacker & Hahn (2000), porque tal critério satisfaz nosso propósito: verificar se as regras de projeção e heurísticas são adequadamente aplicadas. Aspectos mais subjetivos de nossa avaliação ficaram subordinados ao segundo teste, descrito adiante (Seção 5.2.2).

A seguir, detalhamos como as medidas e sua síntese correspondente foram efetuadas.

[6] Por se tratar de uma verificação objetiva, essa comparação poderia ter sido realizada de forma automática, por meio de uma aplicação simples. Optamos pela comparação manual por conta do pequeno volume de dados avaliados.

Medidas individuais

A Tabela 1 ilustra a forma como são tratados os dados sob análise. Tomando o *corpus* de referência como base, para cada sentença, A indica o número de relações UNL do código automático que coincide integralmente com as relações UNL desse *corpus* (sucessos), C, o número de relações do *corpus* de referência que não são geradas pelo ConPor (omissões) e B, o número de relações que são geradas pelo sistema, mas que não constam do *corpus* de referência (falhas).

É interessante notar que uma omissão implica a incapacidade do sistema para identificar uma relação semântica, enquanto uma falha implica a geração de uma relação que é diferente de todas aquelas do *corpus* de referência para aquela sentença, mas que foi considerada legítima pelo ConPor. Ambos os escores podem ser atribuídos para um mesmo par de UWs, ou seja, uma relação UNL não coincidente com as do *corpus* de referência pode indicar, ao mesmo tempo, uma omissão e uma falha. Nesses casos, duas situações são possíveis: a asserção indica uma conceitualização UNL ambígua ou ela é, de fato, um erro.

Tabela 1 – Tabela de contingência para a avaliação dos resultados do ConPor

***Corpus* de referência** / **ConPor**	**Número de relações de referência**	**Número de relações excedentes**
Nº relações geradas	A	B
Nº relações omitidas	C	-

Com base nessa tabela, as medidas individuais são calculadas da seguinte forma:

Cobertura de cada sentença UNL: a parcela de relações que foram geradas corretamente, em relação ao total de relações que deveriam ter sido geradas:

$$\text{CoberturaSent} = \frac{A}{A + C}$$

Precisão de cada sentença UNL: a parcela de relações geradas que são corretas, em relação ao total de relações geradas:

$$\text{PrecisãoSent} = \frac{A}{A + B}$$

***F-measure* de cada sentença UNL:** ponderação sobre os valores da cobertura e precisão, para indicar o desempenho global das regras de mapeamento para cada sentença[7]:

$$\text{F\_measureSent} = 2 * \left(\frac{\text{CoberturaSent} * \text{PrecisãoSent}}{\text{CoberturaSent} + \text{PrecisãoSent}} \right)$$

[7] Uma *f-measure* próxima de 1 indica um ótimo desempenho global do sistema, com relação tanto a sua precisão quando a sua cobertura. Como consideramos o mesmo peso para precisão e cobertura no cálculo da *f-measure*, quando essas duas medidas coincidem, a *f-measure* também coincide.

Um exemplo de análise individual

Como um exemplo do cálculo dessas medidas, considere os dados ilustrados na Tabela 2, que mostra as relações do *corpus* de referência e as relações geradas automaticamente para a sentença (2).

(2) Dê sugestões nessa área.

Tabela 2 – Códigos UNL da sentença (2)

Relações do *corpus* de referência	**Relações geradas pelo ConPor**
obj(give.@entry.@imperative, suggestion.@pl)	obj(give.@entry.@imperative, suggestion.@pl)
mod(area, that)	mod(area, that)
scn(suggestion.@pl, area)	plc(suggestion.@pl, area)

Para essa sentença, a única relação UNL não idêntica é considerada tanto uma omissão (o sistema não gerou uma relação do *corpus* de referência, indicada por *scn*) como uma falha (o sistema gerou uma relação que não consta no *corpus* de referência, indicada por *plc*). Os valores obtidos para a cobertura e precisão coincidem (e, conseqüentemente, *f-measure*) e são iguais a 0,67.

Resultados da análise individual

Da mesma forma que no exemplo ilustrado, foram calculadas as três medidas para as 80 sentenças do *corpus* de teste. Dessas, 77 levaram à cobertura e precisão iguais a 1. Das 3 sentenças restantes, duas tiveram cobertura e precisão iguais a 0 e uma – a sentença (2) – teve cobertura e precisão iguais a 0,67.

Os dois casos para os quais as medidas de cobertura e precisão resultam em 0 correspondem a omissões completas do sistema e ocorrem em função da falta de regras de mapeamento adequadas para algumas características semânticas das sentenças do *corpus* de teste. Ambos os casos, ilustrados pelas sentenças (3) e (4), demandam regras de mapeamento que focalizam verbos da classe "processo": "encontrar" e "curtir", respectivamente[8]. No entanto, as únicas regras aplicáveis às estruturas sintáticas correspondentes focalizam verbos de "ação-processo", o que torna evidente que a semântica dessas entradas não é contemplada pelo sistema.

(3) Seu romantismo encontra receptividade expressiva.

(4) Você poderá curtir a tristeza ou os momentos gratificantes.

Os resultados para a sentença (2), que indicam coincidência parcial entre os códigos do *corpus* de teste e de referência, ocorrem em função da ambigüidade na definição das regras de mapeamento, isto é, o sistema não é capaz de diferenciar entre "lugares físicos" (indicados pelo RL *plc*) e "lugares virtuais" (indicados pelo RL *scn*).

Vale notar que também não foi prevista uma regra de mapeamento completa para a estrutura sintática dessa sentença, considerando verbos de "ação-processo". Seu mapeamento foi possível, todavia, a partir da combinação de partes de regras definidas para outras estruturas sintáticas, conforme discutimos em Specia & Rino (2003).

[8] Uma das características que governa a definição das regras de mapeamento é a classe do verbo principal da sentença, segundo o modelo de classificação verbal de Borba (1990). Maiores detalhes podem ser consultados em Specia & Rino (2003).

Médias

Com base nas medidas individuais, calculamos a média de desempenho das regras de mapeamento do ConPor, ainda em função da sua cobertura, precisão e *f-measure*, conforme fórmula abaixo (NS = número de sentenças do *corpus* de teste; NG = número de sentenças para as quais algum código foi gerado):

$$\text{CoberturaMédia} = \frac{\sum_{i=1}^{NS} \text{CoberturaSent}_i}{NS}$$

$$\text{PrecisãoMédia} = \frac{\sum_{i=1}^{NG} \text{PrecisãoSent}_i}{NG}$$

$$\text{F\_measureMédio} = \frac{\sum_{i=1}^{NS} \text{F\_measureSent}_i}{NS}$$

Os resultados médios obtidos são: **CoberturaMédia** = 0,97, **PrecisãoMédia** = 0,99 e ***F_measureMédio*** = 0,97. Esses valores altos se devem a dois motivos principais: (a) as regras de mapeamento apresentam decisões claras, uma vez que modelam apenas construções simples (devido à natureza do domínio e do público-alvo do sistema); e (b) o *corpus* de teste foi construído para verificar o desempenho do mapeamento para um conjunto limitado de fenômenos lingüísticos, contemplando somente estruturas sintáticas que podem ser manipuladas pela gramática do ConPor.

Embora (a) e (b) possam sugerir uma avaliação tendenciosa, é importante notar que os sucessos, as falhas e as omissões foram identificados de maneira *blackbox*, isto é, independentemente de quaisquer limitações nos módulos do ConPor. Vale lembrar, também, que esses escores foram atribuídos considerando-se o *corpus* de referência, o qual foi criado independentemente das decisões do ConPor.

5.2.2 Teste 2 – Avaliação da expressividade das estruturas conceituais

O objetivo deste teste é avaliar a qualidade das estruturas conceituais geradas (independentemente do mecanismo de geração), em termos do seu poder de expressividade do significado das sentenças originais, isto é, avaliar se tais estruturas expressam adequadamente esse significado. Para tanto, utilizamos a noção de **proximidade semântica**. A proximidade semântica é definida como a proximidade do significado das sentenças originais com relação ao significado das sentenças resultantes da decodificação para o português dos códigos UNL produzidos pelo ConPor. Essa noção é similar à utilizada por Rosé, no entanto, além da preservação da idéia central, ela também considera, explicitamente, a fluência das sentenças

resultantes da codificação. Essa característica é secundária em nossa avaliação, apesar de implicitamente considerada no modo como a proximidade semântica é ranqueada.

Esse teste foi configurado como segue:

Critério: poder de expressividade dos códigos gerados, com relação ao significado das sentenças originais.

Medidas adotadas: média da proximidade semântica do significado associado a cada código UNL gerado pelo ConPor, considerando todos os juízes; média geral, das notas de todos os juízes para todas as sentenças.

Método: primeiramente, os códigos UNL gerados automaticamente pelo ConPor foram decodificados por um especialista em UNL (que não estava envolvido na construção do corpus de teste e, portanto, não teve acesso às sentenças originais)[9]. Em seguida, as sentenças resultantes da decodificação foram comparadas com as sentenças originais do *corpus* de teste, visando medir a proximidade semântica entre cada par de sentenças correspondentes. Para tanto, os pares de sentenças foram submetidos ao julgamento de 27 indivíduos, falantes nativos do português do Brasil, para a atribuição de uma nota para cada par, conforme ilustra a Tabela 3.

Tabela 3 – Opções de avaliação da proximidade semântica

Opções	**Nota**
Nenhuma (significados muito distantes)	1
Mediana (significados relativamente próximos)	2
Completa (significados completamente próximos ou idênticos)	3

Não foi divulgada aos juízes a origem das sentenças sob avaliação – se autênticas ou geradas a partir dos resultados do ConPor.

Proximidade de cada par de sentenças

49 dos 78 pares de sentenças analisados tinham sentenças idênticas. Por esse motivo, não foram apresentadas aos juízes e foram diretamente avaliados com nota 3. Dos pares restantes, os 2 que não tiveram códigos gerados pelo ConPor, tiveram nota 1 atribuída. Os demais 29 pares foram submetidos ao julgamento humano. Com base nas respostas dos juízes, foi calculada a proximidade semântica de cada par de sentenças, computando a nota de todos os juízes, conforme a fórmula a seguir (NJ = número de juízes):

$$\text{ProxSent}_j = \frac{\sum_{i=1}^{NJ} \text{NotaJuiz}_{ij}}{NJ}$$

O resultado do julgamento da proximidade para os 29 pares de sentença é dado na Tabela 4.

[9] Apesar de dispormos de um recurso para decodificação automática (Nunes et al., 1997), optamos pela decodificação manual desse *corpus*, uma vez que tal recurso precisaria ser adequado (em termos da criação de regras) para a decodificação do *corpus* de teste. Além disso, mesmo com a adequação, provavelmente seria necessária a pós-edição humana sobre as decodificações. Assim, da mesma forma que na criação do *corpus* de referência, lançamos mão de um especialista em UNL para tentar garantir codificações/decodificações apropriadas.

Tabela 4 – Resultados da proximidade dos pares de sentenças

Nota	Percentual de sentenças
1 a 2	13,8%
2,1 a 2,9	75,9%
3	10,3%

Proximidade geral

Com base nos resultados das proximidades médias de cada par de sentenças, calculamos também a proximidade geral, para todas as sentenças e juízes, conforme a fórmula abaixo (NS = número de sentenças do *corpus* de teste). Neste caso, também foram incluídos os 49 pares cuja decodificação é idêntica às sentenças originais, bem como os dois casos para os quais nenhum código UNL foi gerado. A proximidade geral resultante é de **2,77**.

$$\text{ProxGeral} = \frac{\sum_{i=1}^{NS} \text{ProxSent}_i}{NS}$$

Considerando que somente 5% dos 80 pares de sentenças tiveram notas médias abaixo da proximidade mediana, e que a média geral obtida foi 2,77 (máximo = 3), podemos concluir que o ConPor produz resultados semanticamente próximos às sentenças originais e, portanto, com alto grau de expressividade do seu significado. Esses valores altos se devem às mesmas razões do Teste 1. Novamente, é preciso considerar que a avaliação foi realizada intuitivamente por humanos, os quais não tiveram acesso às representações UNL e não tinham conhecimento das estratégias de codificação do ConPor.

Como mencionamos, o Teste 2 é menos rígido que o Teste 1, apesar de mais subjetivo, uma vez que a medida de proximidade semântica depende da intuição e experiência do juiz na identificação dos pares de sentença que, embora de não idênticos, podem ainda ser semanticamente próximos. É interessante notar que o critério analisado nesse teste é similar ao dos procedimentos de avaliação de sistemas de tradução automática (a tarefa geral de avaliação de desempenho de White et al. (2000), por exemplo), em particular, para a tradução por interlíngua. A diferença, aqui, é que as línguas fonte e destino são a mesma (português). Devemos considerar, de qualquer modo, que a tradução interlingual se caracteriza mais como uma paráfrase do que como uma transcrição literal da mensagem original em outra língua. Por essa razão, as mensagens decodificadas não precisam ser necessariamente idênticas às originais, devendo apenas manter o significado aproximado para que a tradução seja considerada satisfatória. Nesse sentido, o Teste 2 permite avaliar o nível de influência de uma codificação parcialmente adequada na utilização efetiva dos códigos, de modo a complementar o Teste 1. Por exemplo, a codificação gerada para a sentença (2), aqui repetida, julgada como parcialmente adequada (precisão = 0,67), não impediu a sua decodificação e compreensão, pois a diferença semântica entre os rótulos de relações utilizadas (*plc* e *scn*) é sutil, pelo menos na língua portuguesa.

(2) Dê sugestões nessa área.

Os casos problemáticos

Independentemente do julgamento para a proximidade de todos os pares de sentenças, procuramos identificar as causas das ocorrências de diferenças nas sentenças decodificadas com relação às originais. Agrupamos, para tanto, essas causas em seis itens, discutidos a seguir:

(a) Diferença na ordem da decodificação dos conceitos em conjunções ou disjunções;
(b) Diferença no uso do atributo *@will*;
(c) Decodificação incorreta do gênero dos substantivos;
(d) Inexistência de ordem pré-estabelecida para as relações UNL;
(e) Ambigüidade do rótulo de relação semântica (RL) *aoj* na UNL; e
(f) Não diferenciação dos gêneros masculino e feminino na UNL.

De um modo geral, essas causas não são diretamente relacionadas ao ConPor. As duas primeiras são decorrentes de variações na interpretação de alguns aspectos da UNL, por parte do especialista que realizou a decodificação, com relação ao ConPor. Diferentes interpretações podem implicar diferentes metodologias de codificação e decodificação. Elas ocorrem porque a especificação da UNL não oferece uma definição clara para esses aspectos. A terceira causa corresponde a erros na decodificação, por parte do especialista. As três últimas causas ocorrem em função de problemas da UNL como linguagem de representação semântica. Alguns exemplos das conseqüências dessas causas são ilustrados a seguir, sendo que a primeira de cada par de sentenças é a original e a segunda, a decodificada a partir do código UNL gerado pelo ConPor.

Os pares de sentenças em (5) e (6) correspondem a exemplos de problemas causados por (a) e (b), respectivamente. Com relação à causa (a), no ConPor, o conceito com *@entry* em uma conjunção ou disjunção (neste caso, na relação *and:01(resource.@entry.@pl, partner.@pl)*) indica o primeiro conceito dessa conjunção/disjunção utilizado na sentença (neste caso, o correspondente a "recursos"), ao contrário da decodificação do especialista. Com relação à causa (b), no ConPor o atributo *@will* em um conceito (neste caso, o conceito *heat.@entry.@will* correspondente à "aquecer") indica futuro imediato e não futuro do presente, com o auxiliar "ir", como na decodificação do especialista.

(5) Atraia recursos e associados.
Atraia associados e recursos.

(6) Isso vai aquecer qualquer programa.
Isto irá aquecer qualquer programa.

A causa (c) representa um caso isolado na decodificação, dando origem a um único par de sentenças problemático, ilustrado em (7). Nesse caso, o conceito correspondente a "elevado" está na mesma relação binária UNL (relação de modo) que o conceito de "comunicação" e deveria, portanto, concordar em gênero com esse conceito, mas foi decodificado como sendo do gênero masculino.

(7) Use seu dom de comunicação elevada.
Use seu dom de comunicação elevado.

A mudança de ordem das relações (d) ocorre porque, na UNL, essa ordem é irrelevante, pois se considera que apenas reflete um aspecto sintático da forma das sentenças e, portanto, não tem muita relevância na representação conceitual. Variações na ordem deram

origem a pares de sentenças como os ilustrados em (8) e (9), nas quais o advérbio, que era o primeiro constituinte das sentenças originais, por ter relação direta apenas com o verbo da sentença, foi o último constituinte a ser relacionado pelas heurísticas de relacionamento e, portanto, foi codificado na última relação dessas sentenças[10] e decodificado, posteriormente, no final das sentenças resultantes dos códigos produzidos pelo ConPor.

(8) Muitas vezes você só mostrará alegrias.
Você só mostrará alegrias muitas vezes.

(9) Amanhã a lua em marte aumenta sua coragem.
A lua em marte aumenta sua coragem amanhã.

O RL *aoj* é ambíguo (causa (e)), uma vez que pode representar, de forma indistinta, os diferentes sentidos a que remetem os verbos "ser" e "estar". Por exemplo, ambas as sentenças em (38) são codificadas por meio da relação UNL *aoj(charming.@entry, you)*.

(10) Você está charmoso.
Você é charmoso.

A causa (f) ocorre porque a UNL não diferencia, na sua representação, os conceitos a que remetem os constituintes nominais dos gêneros masculino e feminino. Com isso, pares de sentença como o dado em (11) são representados pelo mesmo conjunto de relações UNL: *gol(be.@entry.@imperative, devoted)* e *to(devoted, beloved.@def)*.

(11) Seja dedicado à amada.
Seja dedicado ao amado.

Os itens (e) e (f), em especial, representam problemas relevantes. Diferentemente da UNL, muitas línguas naturais, além da portuguesa, associam significados diferentes às sentenças como as ilustradas em (10), devido à natureza distinta dos verbos "ser" e "estar". O uso de diferentes gêneros "feminino" e "masculino" em uma sentença também altera o seu significado, em muitas línguas, uma vez que o objeto em foco passa a ser diferente.

Particularmente, para o caso da ambigüidade do RL *aoj*, a especificação da UNL prevê um atributo, *@state*, que pode ser utilizado juntamente com o RL *obj* para indicar estado permanente, causado pela ocorrência de algum evento (geralmente para sentenças com conceitos adjetivais deverbais, como "O vaso está quebrado"). No entanto, esse AL não se aplica a todos os casos de estados, pois não cobre o problema da representação do estado temporário, como ocorre com a primeira sentença em (10). Para indicar estado temporário, a especificação sugere a utilização do RL *aoj*, como fizemos, não permitindo diferenciações entre "ser" e "estar".

Com respeito aos julgamentos para os pares de sentenças, as notas mais baixas dizem respeito a problemas com os verbos "ser" e "estar", que aparecem em 9 pares de sentenças (por exemplo, o par de sentenças em (10), com nota 1,7) e a erros na decodificação de gênero (por exemplo, o par de sentenças em (7), também com nota 1,7). Os melhores casos, isto é, de proximidade bastante acima da mediana, geralmente envolvem a ordem dos componentes da sentença, principalmente nos casos de conjunções (por exemplo, o par de sentenças em (5), com nota 2,8). Os casos médios envolvem algumas ocorrências da ambigüidade do RL *aoj*, de

[10] Para a sentença (8), o código UNL produzido foi *agt(show.@entry.@future, you), obj(show.@entry.@future, happiness.@pl), man(show.@entry.@future, only), man(show.@entry.@future, times), qua(times, many)* e para a sentenca (9), *agt(increase.@entry, moon.@def), plc(moon.@def, mars), obj(increase.@entry, courage), pos(courage, you), tim(increase.@entry, tomorrow)*.

diferenças na ordem de advérbios (por exemplo, o par de sentenças em (9), com nota 2,4 e em (8), com nota 2,2) e de diferença no gênero dos componentes (por exemplo, o par de sentenças em (11), com nota 2,3).

A confiabilidade do teste

Como o julgamento de proximidade semântica é uma medida subjetiva, procuramos considerar o grau de concordância entre os 27 juízes, pois uma concordância indica um nível de confiabilidade suspeito ao teste, podendo invalidá-lo. Para medir esse grau de concordância, utilizamos o índice Kappa (Siegel & Castellan, 1977).

Para calcular esse índice, os resultados fornecidos pelos juízes humanos devem ser distribuídos em uma tabela, de modo a tornar possível extrair os dados sobre a concordância. Neste trabalho, tal tabela foi personalizada, a partir da sugerida pelos criadores do Kappa, para calcular a concordância do segundo teste. Para exemplificar esse processo, na Tabela 5 ilustramos parte da distribuição dos dados e dos resultados obtidos.

Tabela 5 – Distribuição de alguns resultados da avaliação para o cálculo do Kappa

	Classes do julgamento			
Sentenças	**1**	**2**	**3**	**S**
1	0	0	27	1
2	0	0	27	1
3	0	0	27	1
4	8	16	3	0,43
5	0	1	26	0,92
...				
79	0	0	27	1
80	10	16	1	0,47
N = 78	C1 = 81	C2 = 224	C3 = 478	Z = 17,42

As linhas da Tabela 5 indicam todos os itens sendo julgados, neste caso, as sentenças. As três colunas de classes indicam as notas que podem ser atribuídas pelos juízes. Na intersecção de linhas e colunas, para cada sentença é informado o número de juízes que escolheu cada classe. Com base nesses valores da Tabela 5 são levantadas as seguintes informações para o computo do índice Kappa:

J: Número de juízes (27);

N: Número de itens sendo julgados (29);

C1: Total de julgamentos para a classe 1 (81, ou seja, 81 juízes votaram na classe C1, considerando todas as sentenças);

C2: Total de julgamentos para a classe 2 (224);

C3: Total de julgamentos para a classe 3 (478);

S_i: Concordância média da sentença *i*, dada pela fórmula:

$$S_i = \frac{1}{J*(J-1)} * (\sum_{j=1}^{m} n_{ij}(n_{ij}-1))$$

Sendo m = número de classes e n = quantidade de juízes que escolheu a classe *j* para a sentença *i*. Por exemplo, para a quarta sentença da Tabela 5, S seria calculado por essa fórmula instanciada com os valores a seguir, resultando em 0,43:

$$S_4 = \frac{1}{27*(27-1)} * ((8*7) + (16*15) + (3*2))$$

Z : Soma das concordâncias médias de todas as sentenças (17,42), ou seja:

$$Z = \sum_{i=1}^{N} S_i$$

Com base nessas informações, o índice Kappa é calculado por:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Sendo:

P(A): a proporção de vezes que os juízes concordam (0,6):

$$P(A) = \frac{Z}{N}$$

P(E): a proporção de vezes que se espera que os juízes concordem por acaso (0,46):

$$P(E) = \left(\frac{C1}{N*J}\right)^2 + \left(\frac{C2}{N*J}\right)^2 + \left(\frac{C3}{N*J}\right)^2$$

Quanto maior o valor de K, maior a concordância, sendo que K = 1 indica que há uma concordância completa entre os juízes, e K = 0 que não há outra concordância exceto a esperada por acaso. Para o segundo teste da avaliação do ConPor, o valor obtido para Kappa foi **0,25**.

O Kappa não é uma medida específica para o PLN e valores de referência dependem, entre outras coisas, da área de aplicação do índice. Alguns autores, como Carletta et al. (1997), estabeleceram limiares de concordância para suas aplicações (nesse caso, de avaliação de esquemas de codificação de estruturas de diálogo), definindo que o valor 0,8 indica um bom grau de confiabilidade nos resultados, e que valores no intervalo entre 0,68 e 0,8 indicam que a avaliação é suficientemente conclusiva para que algumas inferências possam ser realizadas. Mani (2001), da área de sumarização automática, por sua vez, apenas relata o valor obtido para o índice em um teste (0,24) e discute a dificuldade de identificar se a baixa concordância é devida às diferenças na interpretação do critério em julgamento (aceitabilidade) ou a interpretações similares com julgamentos diferentes. Para a área médica, segundo indica Carletta et al. (1997), valores de kappa entre 0,21 e 0,4 já são considerados aceitáveis para garantir a confiabilidade da avaliação.

Para a avaliação de interpretadores semânticos, em particular, não há ainda limiares ou intervalos de valores "aceitáveis" de confiabilidade. Assim, o valor de Kappa obtido para o segundo teste da avaliação do ConPor, apesar de ser distante de 1, não implica, necessariamente, a baixa confiabilidade do teste, uma vez que pode refletir a discordância natural em função da grande quantidade de juízes e de itens sendo julgados, por exemplo.

6 Considerações finais e trabalhos futuros

Neste relatório descrevemos o processo de avaliação do sistema ConPor, realizado por meio de dois testes, ambos configurando uma avaliação intrínseca autônoma, do tipo

blackbox, baseada em um *corpus* de teste. O Teste1 visa avaliar a qualidade das regras de mapeamento sintático-conceitual e o Teste 2, o poder de expressividade das estruturas conceituais geradas, com relação ao significado das sentenças originais.

O Teste 1 é bastante estrito, uma vez que somente relações UNL idênticas são consideradas "corretas". Apesar de satisfazer propósito de verificar o desempenho das regras de mapeamento, esse teste não garante que as estruturas conceituais geradas pelo ConPor apresentem o mesmo (ou próximo) significado que as sentenças originais. É possível que um mapeamento "correto" indicado por esse teste resulte em uma proximidade semântica insatisfatória. Por outro lado, um mapeamento (parcialmente) incorreto causado, por exemplo, pela impossibilidade do ConPor em reconhecer certas especificidades semânticas, não necessariamente implica sentenças com significado distante. Não existe, portanto, implicação direta entre os valores do Teste 1 e do Teste 2. Dessa forma, a proximidade semântica complementa as medidas tradicionais da avaliação da interpretação semântica (cobertura e precisão), contemplados, aqui, pelo Teste 1. Mais importante, o Teste 2 mostra que a utilização de apenas essas medidas não permite avaliar o principal objetivo da tarefa de interpretação semântica: garantir que as estruturas conceituais produzidas representem adequadamente o significado das sentenças originais. Apesar de Rosé (2000) já ter demonstrado a preocupação com esse critério, sua avaliação difere da do ConPor, uma vez que seu escore "perfeito" inclui tanto a fluência da sentença decodificada quanto o seu grau de comunicação da idéia original. No entanto, nem sempre uma sentença não fluente implica a falta de proximidade semântica. Assim, no ConPor a fluência foi considerada um aspecto secundária com relação à correspondência de significado. Por essas razões, consideramos que o Teste 2 representa uma contribuição para a avaliação de interpretadores semânticos.

Com relação aos resultados do Teste 1, os problemas encontrados se devem à ambigüidade na definição das regras de projeção e à falta de tratamento para determinadas características semânticas das sentenças avaliadas. Já no caso do Teste 2, os principais problemas ocorrem em função de deficiências da linguagem UNL, em particular, a falta de mecanismos para diferenciar a representação dos sentidos distintos a que remetem os verbos "ser" e "estar" e para diferenciar a representação entre os conceitos masculinos e femininos a que remetem os constituintes nominais.

A versão atual do ConPor é limitada, pois focaliza apenas o mapeamento sintático-conceitual. Por isso, ela considera somente uma estrutura sintática como a representação adequada de uma sentença, ignorando a ocorrência natural de ambigüidades sintáticas. Além disso, ela não permite a ambigüidade lexical, o que resulta em codificações bastante diretas. Essas limitações certamente não condizem com as variações naturais na interpretação irrestrita da LN. Extensões dessa versão, de modo a contemplar o tratamento a ambos os problemas, devem ser consideradas em breve. Apesar dessas limitações, o processo descrito neste relatório permite constatar a dificuldade na especificação de procedimentos para analisar as estruturas conceituais *per se*. Além disso, as métricas apresentadas poderiam ser igualmente aplicáveis a abordagens mais refinadas de interpretação semântica, que contemplassem o tratamento às limitações citadas, e permitissem também explorar conjuntos maiores de fenômenos lingüísticos, incluindo sentenças mais longas e complexas.

Como extensões para o processo de avaliação, destacamos a possibilidade de uma avaliação extrínseca (e, possivelmente, comparativa). Outra possibilidade seria uma avaliação *glassbox*, considerando isoladamente os três principais aspectos do sistema: (a) a atribuição dos papéis semânticos aos constituintes da sentença; (b) a seleção dos conceitos, no Léxico Enriquecido, para cada constituinte; e (c) a escolha das relações semânticas entre os pares de conceitos.

Referências Bibliográficas

Bean, C.; Rindflesch, T.C.; Sneiderman, C.A. (1998). Automatic Semantic Interpretation of Anatomic Relationships in Clinical Text. In *Proceedings of the AMIA Annual Fall Symposium*, pp. 897-901. Orlando, Florida.

Black, E.; Abney, S.; Flickenger, D.; Gdaniec, C.; Grishman, R.; Harrison, P.; Hindle, D.; Ingria, R.; Jelinek, F.; Klavans, J.; Liberman, M.; Marcus, M.; Roukos, S.; Santorini, B.; Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. Pacific Grove, California.

Borba, F.S. (1990). *Dicionário gramatical de verbos do português contemporâneo do Brasil*. Fundação Editora Unesp, São Paulo.

Carletta, J.; Isard, A.; Isard, S.; Kowtko, J.C.; Doherty-Sneddon, G.; Anderson, A.H. (1997). The Reliability of a Dialogue Structure Coding Scheme. In *Computational Linguistics*, 23(1), pp. 13-32.

Hirschman, L.; Thompson, H.S. (1996). Overview of Evaluation in Speech and Natural Language Processing. In R. A Cole; J. Mariani; H. Uszkoreit; A. Zaenen and V. Zue (eds), *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge.

Krenn, B.; Samuelsson, C. (1997). *The Linguist's Guide to Statistics: Don't Panic*. Compendium for a course in Statistical Approaches in Computational Linguistics, University of the Saarla. Disponível em http://www.coli.uni-sb.de/~krenn/edu.html/

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company, Philadelphia.

Marcus, M.; Kim, G.; Marcinkiewicz, M.A.; MacIntyre, R.; Bies, A; Ferguson, M.; Katz, K.; Schasberger, B. (1994). The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the Human Language Technology Workshop*, Morgan Kaufmann, San Francisco.

Moore, R.C. (1994). Semantic evaluation for spoken-language systems. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Princeton, New Jersey.

MUC-7 (1998). *Proceedings of the 7th Message Understanding System Evaluation and Message Understanding Conference*. Morgan Kaufman, Washington. Disponível em http://www.saic.com.

Nunes, M.G.V., Vieira, F.M.C., Zavaglia, C., Sossolote, C.R.C., Hernandez, J. (1996). A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos, TR 42 (Relatórios Técnicos do ICMC), USP, São Carlos.

Palmer, M.S.; Finin, T. (1990). *Workshop on the Evaluation of Natural Language Processing Systems*. Unisys Center for Advanced Information Technology, Pennsylvania.

Poesio, M.; Vieira, R. (1998). A *corpus*-based investigation of definite description use. In *Computational Linguistics*, 24(2), pp. 183-216.

Rosé, C. P. (2000). A Framework for Robust Semantic Interpretation. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 311-318. Seattle, WA.

Romacker, M.; Hahn, U. (2000). An Empirical Assessment of Semantic Interpretation. In *Proceedings of the 6th Applied Natural Language Processing Conference & the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 327-334. Seattle, Washington.

Siegel, S.; Castellan Junior, N.J. (1988). *Nonparametric Statistics for Behavioral Sciences*. MacGraw-Hill, New York.

Sparck-Jones, K.; Galliers, J.R. (1996). *Evaluating Natural Language Processing Systems: an Analysis and Review*. Springer-Verlag.

Specia, L.; Rino, L.H.M. (2002a). *ConPor: um gerador de estruturas conceituais UNL*. Série de Relatórios Técnicos do NILC, NILC-TR-02-15. São Carlos, Novembro, 40p.

Specia, L.; Rino, L.H.M. (2002b). *O desenvolvimento de um léxico para a geração de estruturas conceituais UNL*. Série de Relatórios Técnicos do NILC, NILC-TR-02-14. São Carlos, Setembro, 25p.

Specia, L.; Rino, L.H.M. (2003). *A generalização do sistema ConPor*. Série de Relatórios Técnicos do NILC, NILC-TR-03-01. São Carlos, Janeiro, 34p.

Teufel, S.; Moens, M. (1999). Argumentative classification of extracted sentences as a first step towards flexible abstracting. In Inderjeet Mani and Mark T. Maybory (eds.), *Advances in Automatic Text Summarization*. The MIT Press, Cambridge.

UNL (2001). *The Universal Networking Language (UNL) Specifications*. UNU/IAS/UNL Center, Tokyo. Disponível em http://www.unl.ias.unu.edu.

White, J.; Doyon, J.; Talbott, S. (2000). Task tolerance of MT output in Integrated Text Processes. In *Proceedings of the Embedded MT Systems Workshop*, pp. 9-16. Seattle, WA.