Estudo e Desenvolvimento de Modelos Híbridos de Geração de Linguagem Natural

Ivandré Paraboni

Escola de Artes, Ciências e Humanidades (EACH)
Universidade de São Paulo (USP Leste)

Projeto FAPESP nro. 2009/08499-9.

Resumo:

Sistemas de geração de linguagem natural (GLN) – que produzem texto a partir de dados não-linguísticos - possuem uma ampla gama de aplicações em visualização textual de conteúdos complexos e/ou em grandes volumes. Técnicas de GLN desempenham também um papel fundamental em muitas outras aplicações de processamento de linguagem natural como tradução automática, sumarização automática de textos, interfaces para a Web semântica e outras. No contexto do processamento do Português, entretanto, a pesquisa em GLN é ainda relativamente pouco explorada no país. Assim, aproveitando-se a experiência adquirida em nível de doutoramento na área de GLN, propomos um primeiro projeto de maior abrangência neste sentido – o estudo e desenvolvimento de modelos híbridos de realização textual para o Português do Brasil – que sirva de base à pesquisa e desenvolvimento de um grande número de aplicações nesta área, e com o objetivo de longo prazo de consolidação desta linha de pesquisa na instituição na qual está inserido.

1. Enunciado do problema

1.1. Introdução

A visualização de conteúdos de bases de dados complexos e/ou em grandes volumes frequentemente é feita com uso de documentos de textos em linguagem natural. Sistemas de Geração de Linguagem Natural (GLN) – que produzem descrições textuais a partir de uma entrada de dados não linguística – são assim empregados quando texto predefinido não é suficiente, ou seja, quando é necessária uma maior variação linguística nos documentos gerados e/ou maior proximidade em relação ao desempenho humano. Aplicações típicas de GLN incluem, por exemplo, a geração de relatórios descritivos do mercado de ações a partir de indicadores financeiros (Reiter & Dale, 2000), boletins de previsão do tempo gerados a partir de dados de satélites (Belz, 2008), diagnósticos médicos produzidos a partir da leitura de sensores de equipamentos hospitalares (Portet et. al., 2009) e outras. Mais recentemente, também o desenvolvimento da Web semântica - e a crescente necessidade de expressar conteúdos 'complexos' em forma textual - forneceu um novo impulso para a pesquisa em GLN, ao lado, é claro, de aplicações mais tradicionais em diversas linhas de pesquisa do Processamento da Linguagem Natural (PLN) como tradução automática, sumarização, diálogos humano-computador etc.

Embora consideravelmente mais recente do que estudos de interpretação da linguagem[1] - e talvez por este motivo ainda pouco explorada pela comunidade científica brasileira - GLN é uma ativa linha de pesquisa dentro da grande área de PLN. O desenvolvimento de sistemas deste tipo é um empreendimento multidisciplinar de grande escala, frequentemente envolvendo conceitos de diversas áreas do conhecimento, tais como Inteligência Artificial, Psicolinguística e outros. De forma simplificada, entretanto, a arquitetura de um sistema de GLN pode ser vista como um *pipeline* de três estágios (cf. Reiter & Dale, 2000):

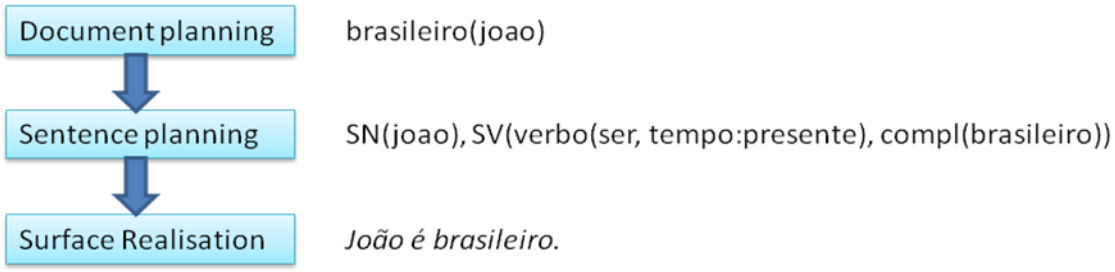

Figura 1 - A arquitetura *pipeline* para GLN proposta em Reiter & Dale (2000)

Tendo como ponto de partida um objetivo de comunicação de alto nível (no exemplo, o objetivo de descrever um objeto do domínio identificado como 'joao'), um sistema de GLN

[1] Estudos na área de interpretação de linguagem são quase tão antigos quanto a própria Ciência da Computação, sendo realizados desde a década de 50. Por outro lado, o estudo da geração de linguagem natural não começou a se desenvolver até o final da década de 70.

constrói gradativamente um plano para representar este objetivo em forma textual (no exemplo, uma sentença em português). Ao longo do processo de geração, o plano inicial sofre uma série de transformações, cada qual constituindo uma forma de representação intermediária do conhecimento comunicado, até o ponto em que o texto em língua natural é obtido. A etapa de *planejamento do documento* ou *macroplanejamento* (Oliveira et. al., 2009) estabelece qual o conteúdo e a estrutura a serem representados no documento final com base nos dados de entrada. A entrada desta etapa é a própria entrada do sistema, ou seja, dados provenientes do domínio da aplicação. O sistema então determina o conteúdo que se deseja comunicar na forma de coleções de mensagens representando o significado pretendido, e.g., na forma de proposições lógicas organizadas em uma estrutura discursiva do tipo RST ou equivalente (cf. Mann & Thompson, 1987)[2]. A etapa seguinte, *planejamento de sentenças* (ou *microplanejamento*) estabelece a organização das mensagens na forma sentencial, ou seja, determina quais construções sintáticas, escolhas lexicais, expressões de referência etc. serão utilizadas para comunicar o conteúdo selecionado (e.g., Lucena & Paraboni, 2008,2008a, 2009,2009a). A entrada desta etapa é um plano do documento (e.g., um conjunto de relações RST e mensagens de conteúdo), e a saída é uma especificação de como o texto será realizado de forma tão independente quanto possível da língua-alvo. Esta representação é finalmente fornecida como entrada para a terceira etapa da geração, a *realização textual*, que utiliza regras gramaticais ou outro recurso compatível para estabelecer mapeamentos entre a especificação abstrata de uma sentença e o texto correspondente em língua natural (Pereira & Paraboni, 2008,2008a; Santos et. al., 2008).

Enquanto as tarefas de macroplanejamento e (em menor grau) microplanejamento são dependentes do domínio da aplicação, a realização textual pode ser abordada de forma relativamente isolada, ou seja, concentrada nos desafios oferecidos pela língua-alvo em questão. Esta independência de domínio - e possibilidade de generalização - tornam a tarefa um objeto especialmente atraente do ponto de vista da pesquisa científica. Além disso, o desenvolvimento de pesquisa enfocando apenas as duas primeiras etapas da arquitetura *pipeline* de GLN tende a enfrentar algumas dificuldades de ordem prática uma vez que o resultado final do sistema (i.e., o texto gerado) não pode ser visualizado sem a disponibilização de um módulo realizador.

Por estas razões, e muito embora a ausência de realização textual certamente não seja um obstáculo à pesquisa em macro/microplanejamento[3], o problema da realização textual em língua Portuguesa nos parece um primeiro passo ideal para a consolidação de uma linha de pesquisa em PLN/GLN nesta instituição, e por este motivo constitui o tema da proposta discutida no restante deste documento. Para uma visão geral da linha de pesquisa em GLN e seus desafios, bem como de técnicas de construção de sistemas deste tipo, sugere-se a leitura de Reiter & Dale (2000) e a extensão de sua arquitetura básica para tratamento de dados da aplicação em Reiter (2007). Uma arquitetura de referência para sistemas de GLN - a

[2] Para resultados da análise RST para o Português do Brasil, ver Pardo & Nunes (2006,2008).

[3] O microplanejamento – que inclui a tarefa de geração de expressões de referência – é de fato um dos temas mais pesquisados em GLN, tendo sido por exemplo objeto de estudo do responsável por esta proposta em diversas ocasiões, e.g., Paraboni & van Deemter (2002,2006), Paraboni et. al. (2006,2007).

especificação RAGS - é descrita em Mellish et. al. (2006), tendo como influência a análise de vários sistemas deste tipo apresentada em Paiva (1998). Finalmente, abordagens baseadas em modelos probabilísticos para conversão de dados em texto incluem, e.g., Belz (2005,2008).

1.2. O problema computacional da realização textual

A realização textual é essencialmente a tarefa computacional de mapear uma representação abstrata do texto – frequentemente já contendo algum tipo de informação estrutural – para uma representação linear em linguagem natural. Sua implementação pode ser dividida em dois componentes relativamente independentes: um componente *tático*, mais dependente da aplicação, encarregado de mapear conceitos de domínio para algum tipo de estrutura abstrata de motivação linguística, e um componente *operacional* encarregado de impor restrições gramaticais e efetuar a linearização de sentenças na língua-alvo. Embora a maioria dos sistemas atuais de alguma forma desempenhe ambas as tarefas, em alguns casos apenas o componente operacional é implementado. Este é o caso por exemplo dos sistemas YAG (McRoy et. al., 2003) e SimpleNLG (Gatt & Reiter, 2009), que partem do princípio de que o mapeamento entre conhecimento semântico e sintático já tenha de alguma forma sido feito, e a preocupação então passa a ser a tarefa de linearização propriamente dita.

Como problema computacional, a realização textual coloca pelo menos três questões de pesquisa que determinam uma ampla gama de estratégias de implementação possíveis: a especificação de entrada, a granularidade da representação gramatical e o paradigma computacional utilizado. Estas questões são discutidas individualmente a seguir no contexto de um projeto de realizador textual para o Português brasileiro.

Uma primeira questão a ser considerada no projeto de um realizador textual é sua *especificação de entrada*: evidentemente, quanto mais próxima esta representação for da forma superficial, menor será o esforço de realização. Ao simplificarmos a tarefa desta forma, entretanto, as decisões de realização estão apenas sendo transferidas para a etapa anterior, isso é, o microplanejamento. Não há de fato um limite claro entre microplanejamento e realização textual, e a decisão sobre o que esperar como entrada para a realização textual é ditada pela aplicação subjacente e o grau de especificidade semântica que esta pode fornecer (Reiter & Dale, 2000). Por exemplo, se a aplicação for capaz de fornecer instruções já detalhadas sobre a lexicalização e a estrutura sintática desejada, então a tarefa de realização textual consistirá basicamente em aplicar um conjunto de regras gramaticais adequado para impor concordância, resolver dependências de longa distância etc. Por outro lado, a inserção deste tipo de processamento linguístico na aplicação pode ter um custo elevado, e que se repetirá a cada novo sistema desenvolvido.

Talvez mais importante do que a questão do custo de uma especificação de entrada detalhada, entretanto, é o fato de que muitas vezes a aplicação simplesmente não necessita, ou não tem condições de exercer total controle sobre a forma linguística final. Nestes casos pode ser mais apropriado o uso de um realizador textual que permita à aplicação fornecer o conhecimento em nível de detalhamento que for tecnicamente viável (ou desejável), ou seja, permitir uma *especificação de entrada neutra e flexível*, possivelmente incompleta do ponto de vista

linguístico, que possa ser adaptada às necessidades de diversos tipos de aplicação com grau de especificidade variado.

Uma segunda questão de projeto, e que é diretamente ligada à definição da especificação de entrada, é a questão da granularidade da *representação gramatical* empregada pelo realizador. Em um extremo, podemos considerar um realizador textual que efetua mapeamentos de conceitos de domínio diretamente para constantes do tipo *string* predefinidas (ou '*canned text*'). Um recurso deste tipo seria de implementação extremamente simples porém sem nenhuma flexibilidade, exigindo manutenção a cada nova variação linguística a ser produzida (além de, é claro, frustrar o próprio objetivo de um sistema de GLN). No outro extremo, podemos considerar um realizador que estabelece mapeamentos entre conceitos de domínio e regras de uma gramática de ampla cobertura da língua-alvo. Esta alternativa proporcionaria máxima flexibilidade e robustez a um grande número de aplicações, mas pode ser de custo elevado, tanto do ponto de vista do desenvolvimento (pois requer o conhecimento de especialistas na língua em questão) quanto da sua utilização prática (já que envolve busca em um espaço de formas linguísticas possíveis). Além disso, observa-se que o uso de uma gramática de ampla cobertura recai no problema anteriormente citado, relacionado ao grau de especificidade de entrada: realizadores com este grau de sofisticação só podem ser plenamente aproveitados se a aplicação puder fornecer uma entrada em nível igualmente detalhado de especificação linguística. Assim, sugere-se que uma solução de custo razoável e de amplo escopo de aplicação prática deveria idealmente implementar uma *representação gramatical intermediária* entre constantes do tipo *string* e um formalismo completo.

Finalmente, quanto ao *paradigma computacional* a ser utilizado, observa-se em Belz (2005) que a maioria dos sistemas de GLN existentes têm sido desenvolvido através de codificação de regras de geração (frequentemente do tipo *if-then*) provenientes de uma cuidadosa análise de requisitos manual baseada em corpus, e com o auxílio de especialistas do domínio. Como as regras que expressam o conteúdo tendem a ser diferentes em cada domínio, na prática o potencial de reutilização de componentes de GLN é mínimo. Neste sentido, a introdução de métodos empíricos aplicados à GLN foi um avanço muito significativo: em anos recentes, o conhecimento proveniente de corpus passou a ser utilizado não apenas na análise de requisitos (como originalmente sugerido em Reiter & Dale, 2000), mas também na tomada de decisões que guiam o próprio processo de GLN (e.g., Pan & Shaw, 2004; Marciniak & Strube 2004,2005; Zhong & Stent, 2005; Belz, 2008), motivando assim um grande número de abordagens baseadas em aprendizagem de máquina e, de forma mais expressiva, modelos estatísticos de língua.

Métodos puramente estatísticos obviamente também possuem limitações. Por exemplo, a abordagem em Belz (2008) considera até 10^{40} alternativas de realização *de cada sentença a ser gerada*, o que pode representar um desempenho computacional insatisfatório para muitas aplicações práticas. Além disso, métodos estatísticos em geral apresentam conhecidas dificuldades de tratamento de dependências de longa distância e, por definição, tendem a favorecer alternativas de realização textual mais breves: por exemplo, em um modelo baseado em *n-gramas* a descrição "o homem de preto" não pode ser menos provável do que "o homem

de preto, à esquerda", e assim tende a ser preferida, o que é claramente inadequado do ponto de vista de uma aplicação de GLN.

Estas dificuldades no entanto não parecem ter diminuído o interesse por métodos estatísticos, interesse esse que é motivado principalmente pelo baixo custo de desenvolvimento e pelo alto grau de reutilização dos recursos desenvolvidos. Afora a relação custo/benefício, entretanto, é evidente que não há uma razão intrínseca para que uma solução de GLN (ou de outra área relacionada) tenha de ser necessariamente limitada a *apenas um único paradigma*. Em especial, observamos que esforços recentes da comunidade de PLN do Brasil têm disponibilizado uma série de recursos linguístico-computacionais potencialmente úteis à tarefa de realização textual tais como *thesauri* (Maziero et. al., 2008), bases lexicais (Muniz, 2004), ontologias e ferramentas associadas (Zavaglia et. al., 2007; Ribeiro Junior & Vieira, 2008), corpora anotados (Abreu et. al., 2007) e outros, cujo reaproveitamento pode não apenas contornar deficiências de um modelo puramente estatístico de realização textual, mas também levar a proposta de soluções inteiramente novas para o problema. Assim, consideramos que uma estratégia de pesquisa promissora para projeto de um realizador textual deva contemplar um paradigma *híbrido*, ou seja, reduzindo custos de desenvolvimento e aumentando a robustez com uso de métodos empíricos, mas sem deixar de tirar proveito do conhecimento simbólico disponibilizado pelos recursos linguístico-computacionais existentes.

Da mesma forma que em muitas outras linhas de pesquisa relacionadas, consideramos que as possibilidades deste enfoque à realização textual são praticamente ilimitadas, e apresentam uma clara oportunidade de avanço do conhecimento na área, e que não se limita a contribuição para o processamento do Português. Por este motivo, a investigação de modelos híbridos de realização textual constitui a idéia central da presente proposta de pesquisa.

1.3. A pesquisa proposta: visão geral

O objetivo da pesquisa proposta é o estudo e desenvolvimento de modelos *híbridos* – que combinam conhecimento linguístico e modelos estatísticos de língua natural – de realização textual para o Português do Brasil, os quais devem ser suficientemente flexíveis para adaptação ao grau de especificidade semântica que venha a ser disponibilizado por diversas aplicações de GLN e mantendo-se um grau satisfatório de variação linguística de saída.

A construção de um realizador textual deste tipo é, até onde temos conhecimento, uma iniciativa inédita de pesquisa científica, e que por ser parcialmente dependente da língua não pode ser desenvolvida através da mera reprodução de trabalhos já existentes para outros idiomas. Entretanto, cabe destacar que apesar do uso do Português como caso de estudo, as contribuições científicas esperadas – e em especial, a proposta de novos modelos computacionais para o problema em questão – devem ser tais que permitam o avanço do conhecimento da área de GLN em geral, e não apenas do processamento do Português.

Com este projeto espera-se também consolidar a linha de pesquisa em PLN/GLN na instituição de destino, fomentando a pesquisa na área e a formação de alunos de iniciação científica tendo em vista a possível implantação de um programa de pós-graduação na

mesma[4]. Além disso, sendo a realização textual o componente mais "visível" (e potencialmente de maior aplicação imediata) de um sistema de GLN, espera-se que sua disponibilização possa também incentivar outros grupos a desenvolver projetos de pesquisa em GLN no país de forma regular.

Do ponto de vista da aplicação, a disponibilização de um realizador textual para o Português abre oportunidade para a pesquisa e desenvolvimento de um grande número de sistemas de geração automática de textos nessa língua, trazendo como consequência imediata a possibilidade de ampliação do acesso ao conhecimento disponível em bases de dados não linguísticos. De forma geral, qualquer aplicação cujo conhecimento seja expresso formalmente (por exemplo, através de ontologias) pode beneficiar-se do produto desta pesquisa. Isso inclui, por exemplo, aplicações de geração automática de respostas em sistemas de *helpdesk* eletrônico, sistemas de geração de diálogos humano-computador ou entre agentes virtuais, aplicações de visualização de bases de dados complexas e/ou extensas, e o amplo horizonte da visualização de conteúdos da Web semântica, dentre muitas outras.

Finalmente, no âmbito da pesquisa do PLN para o Português, observa-se ainda que a tarefa de realização textual é relevante a um grande número de aplicações da área, incluindo tradução automática[5] (e.g., Aziz et. al., 2008,2009; Nunes, Caseli & Focada, 2008), sumarização de textos (e.g., Balage et. al. 2007; Leite et. al., 2007), simplificação textual (e.g., Aluísio et. al., 2008) e muitas outras. Assim, espera-se que os resultados desta pesquisa possam trazer contribuições também nestas frentes e estabelecer colaborações com os centros de pesquisa que as desenvolvem.

2. Resultados esperados

O principal produto do projeto é um componente de software *realizador textual* que recebe como entrada uma especificação semântica proveniente de uma aplicação subjacente (e com flexibilidade para incluir maior ou menor grau de detalhamento do tipo de realização desejado, tal como lexicalização, escolha de tempo verbal etc.) e produz como saída sentenças em Português que expressam o conteúdo especificado.

A construção de um recurso deste tipo traz uma série de contribuições teóricas (e.g., novos modelos híbridos de realização textual de interesse da área como um todo, e também adaptados às características do Português brasileiro com uso de recursos de PLN existentes) e práticas (e.g., a possibilidade de desenvolvimento de um grande número de aplicações de visualização textual de bases de dados extensas e/ou complexas a partir dele). Tomadas em conjunto, estas contribuições representam um avanço significativo - e possivelmente decisivo - para a consolidação da linha de pesquisa em PLN/GLN na instituição na qual esta proposta está inserida.

[4] Uma proposta de curso de mestrado e doutorado da USP/EACH na área de Sistemas de Informação foi recentemente aprovada pela pró-reitoria de pós-graduação e encontra-se em fase de avaliação pela CAPES.
[5] Uma série de considerações sobre a ainda pouco explorada relação entre sistemas de tradução automática e de GLN é apresentada em Knight (2007).

O principal recurso computacional derivado desta pesquisa, bem como outros subprodutos detalhados na proposta a seguir, serão disponibilizados à comunidade científica em geral através de uma página Web do projeto e outros meios tão logo sejam construídos. Além disso, espera-se que todas as suas etapas de estudo e desenvolvimento sejam amplamente divulgadas na forma de publicações científicas das áreas de PLN/GLN e Inteligência Artificial, tanto em âmbito nacional como internacional.

3. Desafios científicos e tecnológicos

Do ponto de vista da pesquisa em GLN para o Português, ambos os componentes de realização textual (tático e operacional) são artefatos inéditos e necessários para o desenvolvimento de aplicações práticas desta natureza. Em especial, observamos que na realização tática a questão do mapeamento entre conceitos do domínio e estruturas abstratas permanece relativamente pouco explorada pela pesquisa na área, carecendo de um tratamento mais sistemático de questões de padronização e reuso, e representa assim um campo de estudo de interesse para a área de GLN como um todo. Quanto ao componente operacional, cabe ressaltar que embora formalismos treináveis a partir de corpora previamente anotados tenham surgido recentemente, independência de língua ainda é um objetivo distante. Por exemplo, em Marciniak & Strube (2005) um realizador textual é derivado a partir de um corpus anotado em um domínio de sentenças muito simples (informações de localização como "dobre à esquerda", "vá até a rua 24th " etc.) Já em DeVault (2008) é apresentado um gerador de diálogos envolvendo a simulação de um agente que fala inglês não nativo, caso em que questões como fluência e gramaticalidade não são consideradas prioritárias. Observamos assim que estudos deste tipo não substituem a necessidade de desenvolvimento de um recurso próprio e com grau de cobertura suficiente (ou de outra forma facilidades de expansão razoáveis) para que a pesquisa e desenvolvimento de aplicações de GLN em Português sejam de fato viabilizadas.

O desenvolvimento de um modelo de realização textual para o Português brasileiro com as características enunciadas na seção anterior (especificação de entrada flexível, representação gramatical de nível intermediário, e uso de um paradigma computacional híbrido) apresenta um número considerável de desafios a serem vencidos. Estes são discutidos no restante desta seção juntamente com possíveis estratégias de solução a serem investigadas.

3.1. Especificação de entrada

Conforme já citado, um primeiro desafio na pesquisa de modelos de realização textual de aplicação prática é a definição de uma especificação de entrada adequada. Sistemas robustos baseados em gramáticas de ampla cobertura (e.g., Langkilde, 2000; White et. al., 2007) em geral esperam uma especificação de entrada igualmente rica, e que originalmente não faz parte da semântica da maioria das aplicações. De fato, em Callaway (2003) observa-se que o custo envolvido no mapeamento da semântica existente para este tipo de entrada pode ser tão elevado quanto o próprio desenvolvimento de um realizador textual específico para cada domínio. Como forma de contornar algumas destas dificuldades, muitos sistemas optam assim por uma especificação de entrada isenta de conhecimento linguístico, ao preço de ter de

desempenhar inclusive as operações de microplanejamento necessárias (em especial, lexicalização) durante a realização textual. Este é o caso por exemplo dos modelos de realização textual apresentados em Corston-Oliver et. al. (2002), Smets et. al. (2003), Marciniak & Strube (2004,2005) e Zhong & Stent (2005), que esperam como entrada a especificação de formas lógicas ou formalismo equivalente. Em uma abordagem distinta, também o trabalho em DeVault et. al. (2008) considera como entrada apenas uma representação semântica na forma de uma vetor de atributos-valores de atos de fala em formato específico para a aplicação considerada.

Com base nestas considerações, coloca-se assim o desafio de projetar um realizador textual em que a especificação de entrada atenda dois requisitos mínimos:

(i) seja representada através de um formalismo de uso geral e padronizado, facilmente adaptável a diversas aplicações de GLN;

(ii) possa ser fornecida pela aplicação subjacente no nível de detalhamento que for conveniente ou possível, e se necessário completada com recursos fornecidos pelo próprio realizador.

Com relação ao requisito (i), consideramos na presente proposta uma alternativa de projeto parcialmente motivada pelo advento da Web Semântica, o qual tem impulsionado um volume sem precedentes de pesquisa e desenvolvimento de *ontologias* para os mais diversos domínios. Muito além de seu papel original, observamos que ontologias representam uma oportunidade ideal para desenvolvimento e teste de sistemas de GLN (em especial, na área de realização tática), pois constituem o tipo de especificação de entrada formal e adequada para uma ampla gama de aplicações da área. De forma mais específica, consideramos o estudo e aproveitamento de *ontologias associadas a corpora* em um domínio específico para a extração de mapeamentos entre conceitos e formas superficiais integrantes do componente tático de realização textual. Uma abordagem similar é empregada, por exemplo, em trabalhos como Marciniak & Strube (2004,2005) para os idiomas inglês e alemão, e no caso do Português pode fazer uso de recursos validados como a ontologia e conjunto de textos anotados em Zavaglia et. al. (2007), ou técnicas de extração semi-automática de ontologias a partir de textos em língua portuguesa (e.g., Ribeiro Junior & Vieira, 2008).

Com relação ao requisito (ii) acima, é desejável que um realizador de propósito geral tenha capacidade de assumir ou completar a especificação de entrada sempre que necessário, ou seja, permitindo à aplicação tanto especificar a forma como deseja expressar cada conteúdo em detalhes (incluindo escolha de cada palavra exata, tempo verbal etc.) como deixar algumas ou mesmo todas estas decisões a cargo do realizador. Na presente proposta o suporte a tais características é proporcionado pela divisão de tarefas de realização entre componentes tático e operacional, e pela combinação de formalismo gramatical baseado em *templates* e paradigma computacional híbrido discutidos a seguir.

3.2. Representação gramatical

O formalismo gramatical empregado pelo realizador textual determina em grande parte o grau de variação linguística possível na saída e, assim como a questão da especificação de entrada,

também apresenta desafios computacionais consideráveis. Conforme já destacado, gramáticas de ampla cobertura da língua-alvo proporcionam máxima flexibilidade e robustez para a tarefa de realização textual. Exemplos deste tipo de abordagem são encontrados em vários realizadores textuais mais antigos, baseados em modelos gramaticais robustos da língua inglesa e outras, como FUF/SURGE (Elhadad & Robin, 1996), KPML (Bateman, 1997) e NITROGEN / HALOGEN (Langkilde & Knight, 1998; Langkilde, 2000). Nestas abordagens, entretanto, o custo computacional elevado pode por exemplo inviabilizar seu uso em aplicações de GLN de tempo real (e.g., DeVault et. al., 2008). Além disso, realizadores baseados em formalismos gramaticais sofisticados podem ser de difícil adaptação à semântica disponibilizada pela aplicação subjacente, que tipicamente inclui pouco ou nenhum conhecimento linguístico.

Por estes motivos, e embora os benefícios de um formalismo gramatical completo sejam indiscutíveis, na presente proposta consideramos o desafio de contornar suas dificuldades adotando uma forma de representação gramatical que:

(iii) equilibre custo (tanto computacional como de desenvolvimento) e variedade linguística, constituindo uma representação intermediária entre constantes do tipo string e gramáticas de ampla cobertura.

Na presente proposta este requisito será atendido com o desenvolvimento de um realizador textual baseado na técnica de *templates*. Templates podem ser vistos como estruturas contendo campos de valores constantes e outros a serem preenchidos com valores fornecidos pela aplicação subjacente, garantindo a ordem e concordância dos seus componentes sem incorrer as dificuldades de construção e uso de um formalismo gramatical completo (McRoy et. al., 2003; van Deemter et. al., 2005; Brugman et. al., 2009). Em sua forma mais simples, um template pode ser representado, por exemplo, como uma estrutura do tipo

<*objeto.nome* " ser " *objeto.atributo*>

cujos campos podem ser preenchidos com base no conhecimento semântico fornecido pela aplicação para gerar uma sentença como "João é brasileiro". Templates podem também representar estruturas sintáticas mais complexas formando árvores de dependência, e seus campos (ou nós-folha) podem ser preenchidos com o conteúdo de outros templates. Tipicamente, realizadores baseados em templates fazem uso de tipos básicos (e.g., templates para sintagmas nominais etc.) associados diretamente a conceitos de domínio, e que podem ser combinados em templates de níveis superiores (e.g., templates de cláusulas ou sentenças). O resultado é uma organização hierárquica que representa as restrições gramaticais da língua-alvo em uma granularidade maior do que uma gramática tradicional, embora possa também ser combinada com estes formalismos (e.g., Becker, 2002).

O emprego de templates de realização textual reduz o tempo de desenvolvimento (especialmente no caso da extração de templates a partir de corpora), a necessidade de especialistas no idioma e a complexidade computacional, já que a operação de busca em um espaço de soluções linguísticas possíveis (como seria o caso de uma gramática) é substituído por um mecanismo de seleção de templates baseado na semântica de entrada (e.g., através de algoritmos de *hashing*, como em McRoy et. al., 2003). Templates são também uma resposta

natural ao requisito de entrada sub-especificada citado no item (ii) da seção anterior. A aplicação neste caso só precisaria selecionar um template compatível com os conceitos a realizar, sem necessidade de acréscimo de instruções linguísticas detalhadas sobre itens lexicais, concordância de gênero ou número, tempo, modo verbal etc. que podem ser tratadas pelas regras associadas ao próprio template. Por outro lado, a aplicação pode também sobrescrever certos valores *default* do template fornecendo desde *strings* predefinidos até conhecimento linguístico no grau que puder ser disponibilizado.

A crítica mais comum aos sistemas baseados em templates é o grau limitado de variação linguística que estes proporcionam. No entanto, a geração baseada em templates não é necessariamente menos sofisticada que o uso de gramáticas de ampla cobertura: sistemas baseados em templates podem ser não apenas de propósito geral (e.g., McRoy et. al., 2003) como são funcionalmente equivalentes a abordagens ditas 'profundas' de GLN (cf. van Deemter et. al., 2005). Além disso, conforme discutido na próxima seção, o uso de templates combinados a modelos estatísticos de língua pode ampliar significativamente a variação linguística e robustez da técnica original.

Finalmente, destacamos ainda que estruturas do tipo template podem ser extraídas de forma manual ou (neste caso em grande escala) de forma semi-automática a partir de corpora contendo anotações morfossintáticas. Assim, no caso da presente proposta - que considera uma especificação de entrada construída a partir de conceitos de ontologias e corpora associados - o uso de templates pode também tirar proveito de recursos de PLN já existentes para o Português conforme já citado (e.g., Zavaglia et. al., 2007).

3.3. Paradigma computacional

Templates utilizam conhecimento linguístico na forma de regras ou formalismo equivalente para estabelecer concordância entre seus componentes e para impor outras restrições estruturais que se façam necessárias. Por exemplo, no template para "João é brasileiro" da seção anterior o campo (ser) deve ser preenchido com a realização deste verbo em uma forma compatível (em gênero, número etc.) com os demais valores da estrutura. O uso de conhecimento linguístico permite assim a definição de templates mais genéricos e com maior poder de expressão do que *strings* predefinidos, mas também torna estas estruturas progressivamente mais próximas de uma gramática da língua-alvo, o que acarreta maior complexidade de desenvolvimento e reutilização.

Problemas deste tipo não são exclusivos da pesquisa em realização textual, mas da área de GLN como um todo. Em Belz (2008) destaca-se que "As pesquisas em outras áreas do PLN chegaram a um estágio em que espera-se que as ferramentas desenvolvidas sejam genéricas, isto é, com ampla cobertura, potencial de reuso e robustez. A pesquisa em GLN, entretanto, ainda falha nos três quesitos." Neste sentido, um avanço significativo foi a explosão em anos recentes do uso de técnicas *estatísticas* de GLN, as quais separam o espaço de geração (possíveis entradas e/ou saídas) do mecanismo de controle de decisões sobre o que produzir: são as chamadas abordagens 'gerar e selecionar' ou 'geração em 2 estágios', introduzidas em Langkilde & Knight (1998) e expandidas na família de sistemas HALOGEN (Langkilde, 2000; Langkilde-Geary, 2002) e outros (Oh & Rudnicky, 2000; Ratnaparkhi, 2000, Varges,

2006). Sistemas deste tipo realizam um processamento predominantemente simbólico para realizar *geração permissiva*, ou seja, produzindo um grande número de soluções possíveis (e.g., a partir de um template básico), incluindo até mesmo alternativas não-gramaticais, e que são filtradas com uso de modelos estatísticos de língua em uma fase de *seleção estatística*. A técnica em 2 estágios permite o desenvolvimento de aplicações de GLN treináveis a partir de corpora e fazendo uso de um volume reduzido de conhecimento linguístico. Técnicas semelhantes também são aplicados à lexicalização (Bangalore & Rambow, 2000) e estruturação sintática (Bangalore & Rambow, 2000a), entre outros.

Modelos mais recentes (e.g., Belz, 2005) ampliam o uso de técnicas estatísticas a outros aspectos da tarefa de GLN, mas em geral o volume de geração de alternativas permanece intenso e de alto custo computacional, além de não ser linguisticamente informado. No caso da presente proposta, observamos que, mesmo com estas limitações, a ausência de recursos linguístico-computacionais para o Português do Brasil teria sido, no passado, uma justificativa suficiente para o uso de modelos puramente estatísticos de GLN. No entanto, este quadro não mais corresponde à realidade. Em anos recentes, foram disponibilizados pela comunidade de PLN do Brasil uma série de recursos potencialmente valiosos para esta pesquisa, e que podem vir a suprir algumas das limitações de métodos puramente estatísticos de realização textual, tais como dicionários (Muniz, 2004), *thesauri* (Maziero et. al., 2008) e outros.

Com base neste novo cenário, uma solução possivelmente ideal de projeto de realizador textual, e alinhada com tendências de pesquisa recente em PLN/GLN, seria tal que:

(iv) possa agregar as vantagens de um paradigma simbólico e estatístico combinados, e tanto quanto possível minimizando as deficiências de cada um.

Para este fim, propomos o estudo de técnicas de geração permissiva e seleção estatística não apenas para filtragem dos resultados da realização textual (como em Langkilde, 2000), mas aplicados também à seleção e preenchimento de templates com uso de conhecimento linguístico proveniente de recursos linguístico-computacionais existentes. Por exemplo, na tarefa de realização tática o mapeamento de conceitos da ontologia para formas superficiais pode usar conhecimento proveniente de um *thesaurus* para propor múltiplas alternativas de lexicalização, as quais podem ser filtradas com uso de um modelo estatístico de língua (e.g., Pereira & Paraboni, 2007) treinado em um corpus do domínio em questão. De forma análoga, a tarefa de realização operacional pode preencher múltiplos templates concorrentes com alternativas de flexão verbal obtidas a partir de um dicionário, e então usar um filtro estatístico para selecionar a alternativa gramaticalmente correta.

Assim como em muitas áreas relacionadas, as oportunidades de experimentação de técnicas de realização textual combinando conhecimento simbólico e modelos estatísticos são praticamente ilimitadas, e sua implementação é perfeitamente possível agora que muitos dos recursos necessários encontram-se disponíveis. Além disso, cabe ressaltar que esta abordagem já se demonstrou tecnicamente viável (ainda que em menor escala) na tarefa de realização textual de descrições definidas em Português (Pereira & Paraboni 2008, 2008a; Santos, Pereira & Paraboni, 2008), em que templates são preenchidos com conhecimento selecionado com uso de modelos estatísticos de língua.

3.4. Plano geral

Nesta seção apresentamos um resumo das atividades de projeto de um realizador textual para o Português do Brasil considerando-se os citados requisitos: (i) especificação de entrada adaptável a um grande número de aplicações; (ii) suporte à sub-especificação; (iii) representação de conhecimento gramatical baseado no uso de templates sintaticamente estruturados e (iv) utilizando-se um modelo computacional híbrido.

O projeto considera a derivação de um realizador textual a partir de uma ontologia e corpus associados, escolha esta parcialmente motivada pelos avanços na área de Web Semântica e tecnologias associadas, que têm gerado um número crescente de aplicações que fazem uso de conhecimento organizado desta forma. Na presente proposta, o uso de conceitos derivados de ontologias como especificação de entrada para a realização textual garante um nível mínimo de padronização e facilidade de adaptação a um grande número de aplicações, conforme requisito (i) acima.

As formas superficiais que realizam cada conceito de domínio (que são obtidos da ontologia) são identificadas no corpus e agrupadas de acordo com as combinações de significados que representam. Por exemplo, todas as sentenças do corpus que descrevem uma ação X sob um objeto Y são agrupadas. A seguir, é computada uma lista de *substrings* (extraídos das formas superficiais) associados a cada conceito, o que corresponde a um conjunto de mapeamentos semântico-sintáticos ou templates de nível inferior (e.g., sintagmas nominais etc.)

Do mesmo corpus são extraídos também os padrões sintáticos sentenciais que correspondem aos templates de níveis superiores (e.g., cláusulas e sentenças). Para cada combinação de conceitos, são identificadas as estruturas sintáticas possíveis, e a mais frequente é selecionada como a base do template. Estas estruturas são então combinadas em uma hierarquia e enriquecidas com regras de concordância de templates, em conformidade com o requisito (ii).

Utilizando-se a estrutura de templates e recursos adicionais de PLN como dicionários e *thesauri*, um mecanismo de geração permissiva é implementado com o objetivo inicial de gerar o maior número possível de realizações textuais para uma dada entrada. A interação entre templates e conhecimento proveniente de bases lexicais e outros recursos implementa o suporte à sub-especificação de entrada do requisito (iii).

Finalmente, o potencial de uso de modelos estatísticos de língua em vários pontos da tomada de decisão do sistema será investigado com o propósito de guiar o processo de geração permissiva, transformando-o em um componente híbrido de GLN de uso prático conforme especificado no requisito (iv).

A ilustração a seguir apresenta um esboço da arquitetura proposta. A porção superior ilustra a constituição da base de templates a partir do corpus e ontologia, e na porção inferior o realizador textual (dividido em módulos de geração permissiva e seleção estatística) exemplifica a transformação de uma entrada altamente sub-especificada (representada apenas por uma forma lógica, abaixo à esquerda) em uma sentença em Português (abaixo, à direita).

Neste exemplo, a fase de geração permissiva faz uso de conhecimento linguístico proveniente de uma base lexical e/ou um thesaurus para selecionar e preencher templates compatíveis com a entrada fornecida, incluída aqui a seleção de um tempo verbal e outras especificações de valores *default*. A fase de seleção estatística (que pode ser aplicada a vários pontos da tomada de decisão, e não apenas ao final da realização textual como a ilustração pode à primeira vista sugerir) faz uso de um modelo estatístico de língua treinado a partir do corpus de trabalho para filtrar a alternativa mais provável dentre as realizações possíveis.

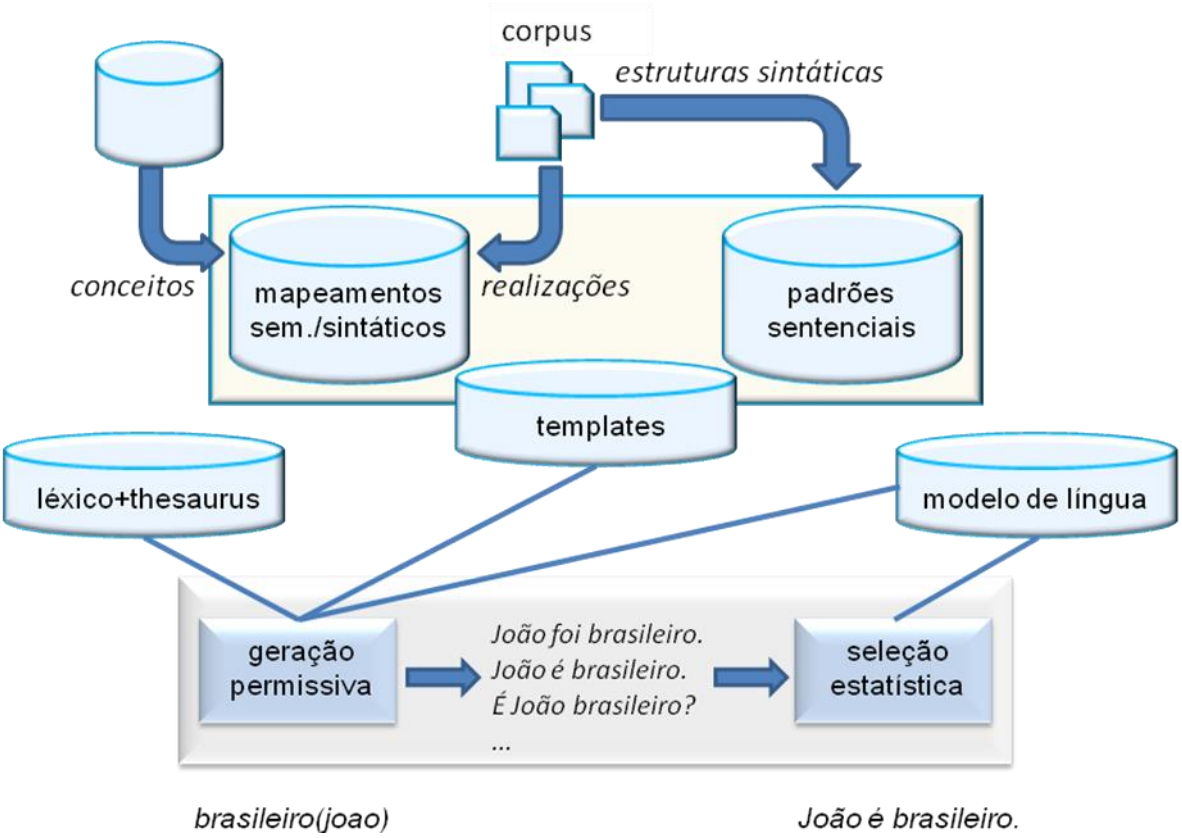

Figura 2 – Um modelo híbrido de realização textual para o Português.

4. Cronograma de Atividades

A seguir apresentamos um cronograma de atividades para desenvolvimento de um realizador textual para o Português do Brasil que admita uma especificação de entrada incompleta e adaptável à semântica disponibilizada pela aplicação subjacente, fazendo uso de templates sintaticamente estruturados e um paradigma computacional híbrido que possa tirar proveito de modelos estatísticos de língua e recursos linguístico-computacionais existentes.

O projeto prevê o início de suas atividades em setembro de 2009 com duração de 24 meses. A seguir relacionamos as principais atividades a serem desenvolvidas e os produtos resultantes de cada uma (destacados em negrito).

1. *Revisão bibliográfica*.
2. *Ambiente de apoio.* Desenvolvimento de recursos básicos a serem utilizados ao longo do projeto. Esta atividade envolve a consolidação de recursos já existentes (módulos de seleção de conteúdo, ferramentas de modelagem de língua etc.) e preparação de um ambiente para desenvolvimento e teste do realizador textual a ser construído.
3. *Preparação do córpus de trabalho*. Seleção de um conjunto de documentos em um domínio específico e identificação dos conceitos relevantes (e.g., com uso de uma ontologia) e das formas textuais que os representam. Com base nestas informações, será selecionado um conjunto de sentenças-alvo a serem geradas (i.e., sentenças que fazem referência aos conceitos em questão). Esta atividade pode ser auxiliada pelo aproveitamento de um recurso já existente como o corpus ECO (Zavaglia et. al., 2007).

   *O resultado desta atividade é um **(a) córpus de sentenças-alvo** a serem geradas.*
4. *Mapeamentos semantico-sintáticos*. Análise do corpus de trabalho para obtenção dos mapeamentos entre conceitos de domínio e realizações superficiais (ou templates de estruturas básicas como sintagmas nominais). Esta etapa corresponde ao desenvolvimento de um módulo tático de realização textual.

   *O resultado desta atividade é um **(b) dicionário de realizações** dos conceitos do domínio.*
5. *Extração de padrões sentenciais*. Extração das estruturas sintáticas existentes no corpus de trabalho, considerando-se também certas variações adicionais que podem não estar presentes no corpus (e.g., uso de voz passiva etc.) mas cuja modelagem possa ser de interesse para uma solução de maior abrangência.

   *O resultado desta atividade é uma **(c) base de estruturas sintáticas** na forma de árvores sintáticas em que certas folhas representam lacunas a serem preenchidas com realizações textuais de conceitos do domínio que representam.*
6. *Construção da base de templates*. Implementação de uma organização hierárquica de templates básicos (os mapeamentos do dicionário (b) acima) e construções sintáticas possíveis (produto (c) acima).

   *O resultado desta atividade é uma **(d) base de templates sintaticamente estruturados**.*

7. *Geração permissiva*. Projeto e implementação de um algoritmo básico para realização textual permissiva e com baixo teor de conhecimento linguístico, que dada uma representação semântica de entrada retorna uma floresta de construções possíveis, incluindo possíveis alternativas não-gramaticais.

   *O resultado desta atividade é um modelo básico de **(e) realizador textual permissivo**.*

8. *Integração de recursos linguístico-computacionais*. Expansão das opções de saída com acréscimo de recursos para geração do maior número possível de alternativas (e.g., obtidas a partir de dicionários de flexões e *thesauri*) e suporte à especificação de entrada incompleta, a serem posteriormente filtradas com uso de modelos estatísticos.

9. *Seleção estatística*. Investigação, implementação e teste de técnicas de uso de modelos estatísticos de língua em diferentes etapas da geração (e.g., lexicalização, concordância entre templates etc.) para obtenção da alternativa ideal (i.e., a mais provável) dentre as múltiplas soluções possíveis. Esta atividade será apoiada por um projeto complementar em nível de Iniciação Científica (em fase de proposta) tratando exclusivamente da experimentação com técnicas estatísticas aplicadas à realização textual.

   *O resultado desta atividade é um modelo **(f) realizador textual híbrido** que é o produto final desta pesquisa.*

10. *Avaliação*. Seleção de uma aplicação exemplo e uso de métricas automáticas como NIST/BLEU (Papineni et. al., 2001) para avaliação objetiva da qualidade textual obtida em relação a um corpus de referência. Estas métricas serão possivelmente complementadas com uso de técnicas de avaliação subjetiva empregando julgadores humanos e outras, conforme discutido em Bangalore et. al. (2000).

11. *Disseminação:* divulgação de resultados em reuniões, eventos científicos e publicações qualificadas da área em todas as etapas do projeto.

O cronograma de execução destas atividades (1-11) e produtos (a-f) indicadores do andamento de cada etapa são apresentados a seguir.

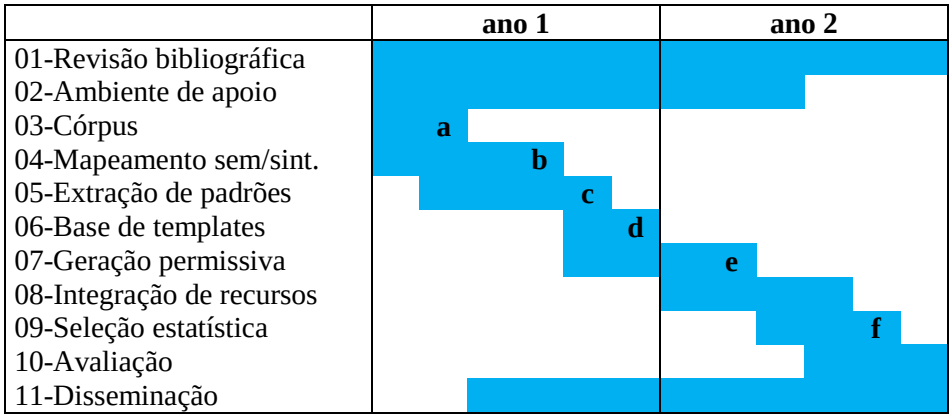

Figura 2 - Cronograma bimestral de atividades propostas (1-11) e seus produtos (a-f)

5. Disseminação e avaliação

O projeto prevê etapas de revisão das atividades de análise e extração de informações de corpus (produtos a,b,c) e validação do resultado final (o realizador textual em suas duas versões representadas pelos produtos e,f) em uma etapa específica de avaliação com uso de métricas objetivas e subjetivas conforme descrito na atividade 10 do cronograma.

Os recursos computacionais desenvolvidos serão disponibilizados à comunidade científica em geral através de uma página Web do projeto e outros. Além disso, espera-se que todas as suas etapas de estudo e desenvolvimento sejam amplamente divulgadas na forma de publicações científicas das áreas de PLN/GLN e Inteligência Artificial em âmbito nacional e internacional.

6. Outros apoios

No momento não há apoio financeiro específico para este projeto. No entanto, a motivação para esta proposta partiu de dois projetos relacionados na área de PLN (um recentemente concluído, e outro com conclusão prevista para outubro próximo) que agregaram alguns recursos de apoio à pesquisa conforme detalhado na documentação anexa:

- *Construção de um córpus paralelo com informações de correferência para tradução automática* (*Novembro 2006- outubro 2008)* Projeto de pesquisa individual FAPESP nro. 2006/03941-7.
- *Resolução e Interpretação de Expressões de Referência na Geração Automática de Textos* (*Novembro 2007 – outubro 2009)* Projeto de pesquisa individual CNPq (edital Universal) nro. 484015/2007-9.

Além destes, os seguintes projetos foram beneficiados com bolsas de Iniciação Científica sob responsabilidade do autor desta proposta:

Bolsas de Iniciação Científica (concluídas):

Ramon Ré Moya Cuevas. *Aperfeiçoamento de Técnicas Estatísticas de Tradução Automática pelo Tratamento de Correferências*. USP / EACH. Bolsista PIBIC (USP). Conclusão: novembro 2007.

Daniel Bastos Pereira. *Avaliação Automática de Textos em Sistemas de Tradução Estatística*. USP / EACH. Bolsista do programa USP "Ensinar com Pesquisa". Conclusão: janeiro 2008.

Wilker Ferreira Aziz. *Investigação de Técnicas de Alinhamento Textual para a Tradução Automática Estatística*. USP / ICMC. Bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo. Conclusão: fevereiro 2008.

Diego Jesus de Lucena *Investigação de Técnicas de Aprendizagem de Máquina para o Processamento de Línguas Naturais*. USP / EACH. Bolsista CNPq. Conclusão: julho 2008.

Ramon Ré Moya Cuevas *Investigação de Técnicas Computacionais de Resolução de Referências Pronominais*. USP / EACH. Bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo. Conclusão: janeiro 2009.

Daniel Bastos Pereira *Desenvolvimento e Avaliação de Modelos Estatísticos de Língua*. USP / EACH. Bolsista do programa USP "Ensinar com Pesquisa". Conclusão: fevereiro 2009.

Wilker Ferreira Aziz *Desenvolvimento de um Protótipo de Sistema de Tradução Automática Estatística*. USP / ICMC. Bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo. Conclusão: março 2009.

Bolsas de Iniciação Científica (vigentes):

Diego Jesus de Lucena *Estudo de Métodos Empíricos no Processamento de Linguagem Natural*. USP / EACH. Bolsista CNPq.

Rafael Lage de Oliveira *Planejamento de Documentos em Aplicações de Geração Automática de Linguagem Natural*. USP / EACH. Fundação de Amparo à Pesquisa do Estado de São Paulo.

Eder Miranda de Novais. *Seleção de Conteúdo para Geração Automática de Textos em Linguagem Natural*. USP / EACH. Bolsista CNPq (Iniciação em desenvolvimento tecnológico e inovação).

Roberto Paulo Andrioli de Araújo. *Geração Automática de Estruturas Discursivas*. USP / EACH. Bolsista do programa USP "Ensinar com Pesquisa".

Thiago Dias Tadeu. *Planejamento de Documentos na Geração Automática de Textos: Preparação para Realização Textual*. USP / EACH. Bolsista CNPq / PIBIC.

7. Bibliografia

Abreu, Sandra Collovini de, T. I. Carbonel, J. C. B. Coelho, J. T. Fuchs, L. H. M. Rino e R. Vieira (2007) *Summ-it: um corpus anotado com informações discursivas visando à sumarização automática*. TIL 2007 Workshop de Tecnologia da Informação e da Linguagem Humana.

Aluísio, S. M. , L. Specia, T. A. S. Pardo, E. Maziero e R. P. M. Fortes (2008) *Towards Brazilian Portuguese Automatic Text Simplification Systems*. The ACM Symposium on Document Engineering, São Paulo, pp. 240-248.

Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo e Ivandré Paraboni (2008) *An Experiment in Portuguese-Spanish Statistical Machine Translation*. 19th Brazilian Symposium on Artificial Intelligence (SBIA-2008). LNAI vol. 5249, pp. 248-257. Springer-Verlag Berlin Heidelberg.

Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo e Ivandré Paraboni (2009) *Statistical Phrase-based Machine Translation: Experiments with Brazilian Portuguese*. XXIX CSBC / VII Encontro Nacional de Inteligência Artificial (ENIA-2009). 20-24 de julho, Bento Gonçalves, RS.

Balage Filho, P.P, Thiago Alexandre Salgueiro Pardo e Maria das Graças Volpe Nunes (2007) *Summarizing Scientific Texts: Experiments with Extractive Summarizers*. 7th International Conference on Intelligent Systems Design and Applications, Rio de Janeiro, pp. 520-524.

Bangalore, S. e O. Rambow (2000) *Corpus-based lexical choice in natural language generation*. 38th Meeting of the Association for Computational Linguistics (ACL'00), Hong Kong, pp. 464-471.

Bangalore, S. e O. Rambow (2000a) *Exploiting a probabilistic hierarchical model for generation*. 18th International Conference on Computational Linguistics (COLING '00), pp. 42–48.

Bangalore, S., O. Rambow e S. Whittaker (2000) *Evaluation metrics for generation*. 1st International Conference on Natural Language Generation (INLG '00), pp.1-8.

Bateman, J. A. (1997) *Enabling technology for multilingual natural language generation: the KPML development environment.* Natural Language Engineering 3(1), pp.15-55.

Becker, Tilman (2002) *Practial, Template-Based Natural Language Generation with TAG*. 6th Intl. Workshop on Tree Adjoning Grammars and Related Frameworks (TAG+6). Veneza, pp.101-104.

Belz, Anja (2005) *Statistical Generation: Three Methods Compared and Evaluated*. Proceedings of the 10th European Workshop on Natural Language Generation (ENLG'05).

Belz, Anja (2008) *Automatic Generation of Weather Forecast Texts using Comprehensive Probabilistic Generation-Space Models.* Natural Language Engineering 14 (4), pp. 431-455.

Brugman, I., Mariët Theune, Emiel Krahmer e Jette Viethen (2009) *Realizing the Costs: Template-Based Surface Realisation in the GRAPH Approach to Referring Expressions Generation.* 12th European ws. on Natural Language Generation (EACL/ENLG 2009) Atenas, Grécia, pp. 183-184.

Callaway, C. B. (2003) *Evaluating coverage for large symbolic NLG grammars.* IJCAI 2003.

Corston-Oliver, S., Michael Gamon, Eric Ringger e Robert Moore (2002) *An overview of Amalgam: A machine-learned generation module*. 2nd Intl. Conference on Natural Language Generation.

DeVault, David, David Traum e Ron Arstein (2008) *Practical Grammar-Based NLG from Examples*. 5th International Natural Language Generation Conference (INLG-2008) Columbus, USA.

Elhadad, M. e J. Robin (1996) *An overview of SURGE: A reusable comprehensive syntactic realization component.* 8th International Natural Language Generation workshop.

Gatt, Albert e Ehud Reiter (2009) *SimpleNLG: A realisation engine for practical applications*. EACL / ENLG-2009.

Knight, Kevin (2007) *Automatic Language Translation Generation Help Needs Badly* (2007) MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UCNLG+MT), pp. 1-4.

Langkilde, Irene and Kevin Knight (1998) *Generation that exploits corpus-based statistical knowledge*. COLING-ACL'98, pp.704-710, Montreal, Canada.

Langkilde, Irene (2000) *Forest-based statistical sentence generation*. 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association of Computational Linguistics (ANLP-NAACL '00), pp. 170–177.

Langkilde-Geary (2002) *An empirical verification of coverage and correctness for a general-purpose sentence generator*. International Natural Language Generation Conference (INLG-2002).

Leite, D. S., Lucia Helena Machado Rino, Thiago Alexandre Salgueiro Pardo e Maria das Graças Volpe Nunes (2007) *Extractive Automatic Summarization: Does more linguistic knowledge make a difference?*. TextGraphs-2 HLT/NAACL Workshop, Rochester.

Lucena, Diego Jesus de e Ivandré Paraboni (2008) *Frequency-based Greedy Attribute Selection for Referring Expressions Generation*. 5th International Natural Language Generation Conference (INLG-2008) Salt Fork, USA..

Lucena, Diego Jesus de e Ivandré Paraboni (2008a) *Combining Frequent and Discriminating Attributes in the Generation of Definite Descriptions*. 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA-2008) LNAI vol. 5290, pp. 252-261. Springer-Verlag Berlin Heidelberg.

Lucena, Diego Jesus de e Ivandré Paraboni (2009) *Improved Frequency-based Greedy Attribute Selection*. 12th European ws. on Natural Language Generation (EACL/ENLG 2009) Atenas, Grécia.

Lucena, Diego Jesus de e Ivandré Paraboni (2009a) *The Design of an Experiment in Anaphora Resolution for Referring Expressions Generation*. Recent Advances in Natural Language Processing (RANLP-2009).

Mann, W. C. and S. A. Thompson (1987) Rhetorical Structure Theory: A Theory of Text Organisation L. Polanyi (ed.) *The Structure of Discourse*. Ablex, Norwood, USA.

Marciniak, T. e M. Strube (2004) *Classification-based generation using TAG*. 3rd Intl.Conference on Natural Language Generation (INLG'04), LNAI vol. 3123, pp.100–109. Springer-Verlag.

Marciniak, T. e M. Strube (2005) *Using an Annotated Corpus As a Knowledge Source For Language Generation*. Corpus Linguistics'05 Workshop Using Corpora for NLG (UNNLG-2005), pp.19-24.

Maziero, Eick G., Thiago A. S. Pardo, Ariani di Felippo e Bento C. Dias-da-Silva (2008) *A Base de Dados Lexical e a Interface Web do TeP 2.0 –Thesaurus Eletrônico para o Português do Brasil*. VI Workshop on Information and Human Language Technology (TIL-2008).

McRoy, Susan, Songsak Channarukul e Syed S. Ali (2003) *An augmented template-based approach to text realization*. Natural Language Engineering 9 (4) pp. 381–420. Cambridge University Press.

Mellish, C. et. al. (2006) *A Reference Architecture for Natural Language Generation Systems*. Natural Language Engineering 12 (1) pp.1–34.

Muniz, M. C. M. (2004) *A construção de recursos linguístico-computacionais para o português do Brasil: o projeto de Unitex-PB*. Dissertação de Mestrado. ICMC / USP São Carlos.

Nunes, Maria das Graças Volpe, Helena de Medeiros Caseli e Mikel Forcada (2008) *Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation*. Machine Translation, v. 20, pp. 227-245.

Oh, A. e A. Rudnicky (2000) *Stochastic language generation for spoken dialogue systems*. ANLP-NAACL 2000 Workshop on Conversational Systems, pp.27–32.

Oliveira, Rafael Lage de, Eder Miranda de Novais, Roberto Paulo Andrioli de Araújo e Ivandré Paraboni (2009) *A Classification-driven Approach to Document Planning*. Recent Advances in Natural Language Processing (RANLP-2009).

Paiva, Daniel (1998) *A Survey of Applied Natural Language Generation Systems*. ITRI Technical Report ITRI-98-03. University of Brighton, United Kingdom.

Pan, Shimel e James Shaw (2004) *SEGUE: A Hybrid Case-Based Surface Natural Language Generator*. 3rd Intl. Conference on Natural Language Generation (INLG'04), LNAI vol. 3123.

Papineni, K., S. Roukos, T. Ward e W.-J. Zhu (2001) *BLEU: A method for automatic evaluation of machine translation*. IBM research report, IBM Research Division.

Paraboni, Ivandré e Kees van Deemter (2002) *Towards the Generation of Document-Deictic References*. In: Information Sharing: Reference and Presupposition in Language Generation and Interpretation. CSLI Publications, Stanford, USA, pp.329-354. ISBN

Paraboni, Ivandré e Kees van Deemter (2006) *Referring via document parts*. 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006) Cidade do México, 19-25 fevereiro. LNCS vol. 3878, pp.299-310. Springer-Verlag Berlin Heidelberg.

Paraboni, Ivandré, Judith Masthoff e Kees van Deemter (2006) *Overspecified reference in hierarchical Domains: measuring the benefits for readers*. 4th International Natural Language Generation Conference (INLG-2006) Sydney, Australia, pp.55-62.

Paraboni, Ivandré, Kees van Deemter e Judith Masthoff (2007) *Generating Referring Expressions: Making Referents Easy to Identify*. Computational Linguistics 33(2), junho, pp. 229-254.

Pardo, Thiago Alexandre Salgueiro and Nunes, Maria das Graças Volpe (2006) *Review and Evaluation of DiZer - an Automatic Discourse Analyzer for Brazilian Portuguese*. 7th Workshop on Computational Processing of Written and Spoken Portuguese.

Pardo, Thiago Alexandre Salgueiro and Nunes, Maria das Graças Volpe (2008) *On the Development and Evaluation of a Brazilian Portuguese Discourse Analiser*. Revista de Informática Teórica e Aplicada, v. XV, p. 43-64.

Pereira, Daniel Bastos e Ivandré Paraboni (2007) *A Language Modelling Tool for Statistical NLP*. 5th Workshop on Information and Human Language Technology (TIL-2007) pp.1679-1688.

Pereira, Daniel Bastos e Ivandré Paraboni (2008) *Statistical Surface Realisation of Portuguese Referring Expressions*. 6th International Conference on Natural Language Processing (GoTAL-2008) Gothenburg, Suécia. LNAI vol. 5221, pp. 383-392. Springer-Verlag Berlin Heidelberg.

Pereira, Daniel Bastos e Ivandré Paraboni (2008a) *From TUNA Attribute Sets to Portuguese Text: a First Report*. 5th Intl. Natural Language Generation Conference (INLG-2008) Salt Fork, USA.

Portet, F., E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer e C. Sykes (2009) *Automatic Generation of Textual Summaries from Neonatal Intensive Care Data*. Artificial Intelligence 173, pp. 789-816

Prasad, Rashmi Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki e Bonnie Webber (2005) *The Penn Discourse TreeBank as a Resource for Natural Language Generation*. Corpus Linguistics'05 Workshop Using Corpora for NLG (UNNLG-2005), pp.25-32.

Ratnaparkhi, A. (2000) *Trainable methods for surface natural language generation*. ANLP-NAACL'00, pp.194–201.

Reiter, Ehud and Robert Dale (2000) *Building natural language generation systems*. Cambridge University Press.

Reiter, Ehud (2007) *An Architecture for Data-to-Text Systems*. Proc. of ENLG-2007, pp. 97-104.

Ribeiro Junior, Luiz Carlos e Renata Vieira (2008) *OntoLP: Engenharia de Ontologias em Língua Portuguesa*. XXVIII Congresso da Sociedade Brasileira de Computação. Porto Alegre: SBC.

Santos, Francis Marques Veras dos, Daniel Bastos Pereira e Ivandré Paraboni (2008) *Rule-based vs. Probabilistic Surface Realisation of Definite Descriptions*. VI Workshop on Information and Human Language Technology (TIL-2008). XIV Brazilian Symposium on Multimedia and the Web.

Smets, Martine, Michael Gamon, Simon Corston-Oliver e Eric Ringger (2003) *French Amalgam: A machine-learned sentence realization system.* TALN-2003, Batz-sur-Mer, 11-14 July.

van Deemter, K., Emiel Krahmer e Mariët Theune (2005) *Real versus template-based NLG: a false opposition?* Computational Linguistics 31(1).

Varges, Sebastian (2006) *Overgeneration and ranking for spoken dialogue systems*. 4th International Natural Language Generation Conference (INLG-2006) Sydney, Australia, pp. 20-22.

White, M., Rajakrishnan Rajkumar and Scott Martin (2007) *Towards Broad Coverage Surface Realization with CCG*. MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UCNLG+MT), pp.22-30.

Zavaglia, Claudia, Leandro Henrique Mendonça de Oliveira, Maria das Graças Volpe Nunes e Sandra Maria Aluísio (2007) *Estrutura Ontológica e Unidades Lexicais: uma aplicação computacional no domínio da Ecologia*. Anais do XXVII Congresso da SBC. Rio de Janeiro, 5-6 julho, pp.1575-1584.

Zhong, Huayan e A. J. Stent (2005) *Building Surface Realizers Automatically from Corpora*. Corpus Linguistics'05 Workshop Using Corpora for NLG (UNNLG-2005), pp.49-54.