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

Especificação Lingüística de um Realizador Superficial Baseado em Decisões de Estilo

Mauricio José Carvalho de Bem
Lucia Helena Machado Rino

NILC-TR-02-17

Setembro, 2002

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

Resumo

Neste relatório apresentamos as características lingüísticas de um realizador superficial baseado em decisões sobre alguns estilos do português, indicando o conteúdo e a função de seus principais repositórios lingüísticos.

Para construir tais recursos, baseamo-nos em sentenças selecionadas de um corpus, que passaram a ser nosso conjunto de exemplos básicos. Ilustrações e descrições sobre a forma como o realizador foi especificado são exploradas aqui.

Este trabalho conta com o
apoio financeiro do CNPq (PIBIC)

Índice

Índice de Figuras

Índice de Tabelas

1 Introdução

Neste relatório descreveremos a metodologia adotada para a construção dos recursos lingüísticos de um realizador superficial automático cuja principal característica é tomar decisões com base em algumas informações de estilo da língua portuguesa.

A construção de um protótipo desse realizador foi proposta como Projeto PIBIC (Agosto/2001 – Julho/2002) e é baseada na escolha de exemplos típicos, que servem de base para a especificação dos recursos lingüístico-computacionais. Como trabalho de Iniciação Científica, esse protótipo é bastante simples. No entanto, a idéia é que ele possa ser expandido gradualmente, pela incorporação de outras características de estilo.

No que segue, descrevemos brevemente a arquitetura do protótipo (Seção 2), para identificar os recursos lingüísticos cuja especificação é o foco deste relatório. Primeiramente, apresentamos um corpus escolhido como base para a prototipação (Seção 3), o qual permitiu identificar as variações estilísticas a serem consideradas (Seção 4). A partir destas, pudemos especificar o léxico e o subconjunto das regras gramaticais do português, descritos, respectivamente, nas Seções 5 e 6.

2 A arquitetura do realizador superficial

O realizador superficial cuja arquitetura é descrita na Figura 1 considera como entrada uma estrutura conceitual, que deverá ser analisada para determinar possíveis desvios estilísticos. Consideram-se estruturas conceituais como estruturas sintáticas, profundas, correspondente a sentenças originalmente escritas em português. Como a análise sintática (*parsing*) não é parte deste projeto, tais estruturas são construídas semi-automaticamente, fazendo-se uso de um *parser* (detalhes sobre isso serão dados na próxima seção).

A detecção de tais desvios é baseada em um conjunto de regras de estilo, que constitui o recurso 'Regras de estilo' na Figura 1. Identificado o desvio presente na estrutura conceitual, o realizador superficial busca uma regra de mapeamento estilístico (recurso 'Regras de mapeamento estilístico') para, por exemplo, trocar um item lexical que expressa o desvio por outro, determinado pela regra em foco. Nesta etapa, claramente também é utilizado um léxico da língua portuguesa, o qual é importado do léxico do

NILC[1] (Nunes, 1996). Caso o desvio estilístico seja de natureza sintática, a regra de mapeamento indicará quais as regras gramaticais do português que deverão ser empregadas para alterar a estrutura conceitual de entrada, produzindo a nova estrutura conceitual da sentença pretendida. Esta estrutura é, então, linearizada, resultando na sentença de saída, em português, que não mais conterá o desvio apontado no início.

Assim, o realizador tem a função de propor variações estilísticas de uma sentença original, porém, ele não terá a própria sentença como entrada, mas já sua estrutura conceitual resultante da análise sintática da mesma. Essa estrutura é representada como uma estrutura-f (Kaplan & Bresnan, 1982).

Nesse cenário, um usuário pode, p.ex., especificar o grau de formalidade da sentença que o protótipo deve produzir, a partir do qual o sistema deverá eleger tanto as palavras quanto suas estruturas finais. Decisões dessa natureza são interdependentes e deverão ser modeladas de maneira simplificada neste trabalho.

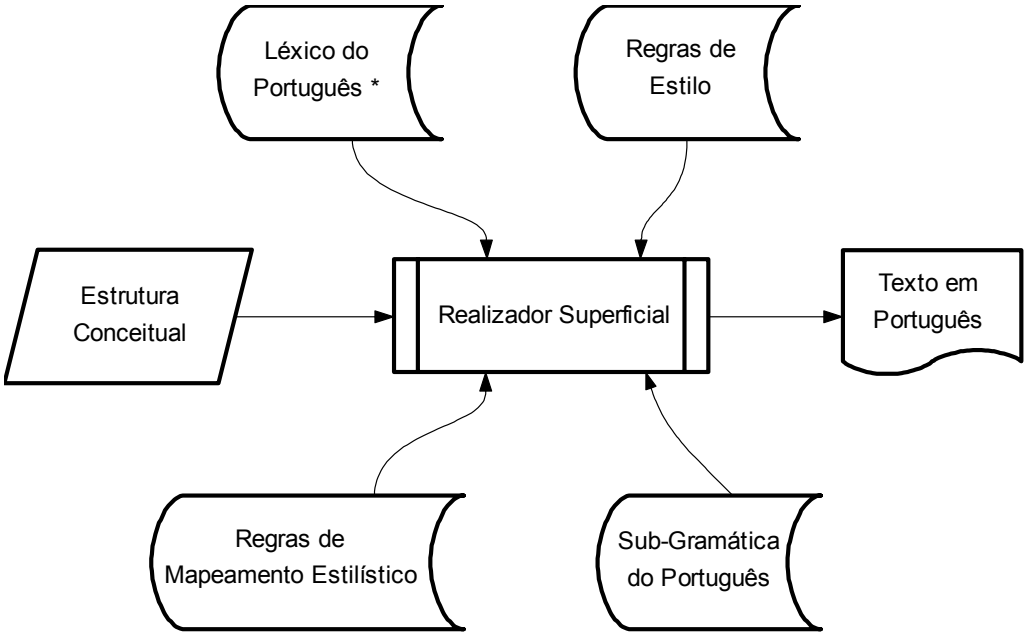

Figura 1 – Arquitetura do realizador superficial

Apresentamos, a seguir, as fases do projeto lingüístico do protótipo, descrevendo as principais decisões de especificação.

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

3 Corpus de prototipação

Para a construção do protótipo, selecionamos somente algumas sentenças do português, a partir das quais os recursos apontados na Figura 1 e o próprio módulo de processamento foram definidos. O léxico foi limitado aos itens lexicais de representação das sentenças consideradas e a eventuais entradas relacionadas às regras de estilo em foco. Analogamente, somente foram consideradas as regras gramaticais do português que tinham relação com os requisitos de realização superficial. É por esse motivo que especificamos o recurso gramatical correspondente como uma sub-gramática, também delimitando o conjunto de regras de estilo e de regras de mapeamento entre a estrutura conceitual de entrada e as possíveis estruturas sentenciais das saídas, as quais deverão apresentar as variações estilísticas apontadas pelas regras de mapeamento.

O corpus escolhido é composto somente por dez sentenças extraídas do corpus corrigido do NILC (Kuhn et al., 2000), com base nos textos de vários cadernos do jornal "Folha de S. Paulo" (TV Folha, Caderno Especial, Veículos, Ilustrada e Esporte). Esse corpus foi escolhido porque supusemos que, por considerar uma ampla audiência e questões atuais, os textos poderiam evidenciar estilos diversos, implicando maior possibilidade de variações estilísticas. Entretanto, tivemos que limitá-lo tanto em número quanto em termos da complexidade das sentenças apresentadas, visando minimizar a dificuldade de simular o processamento automático. A Figura 2 mostra o conjunto de sentenças selecionadas para dar início à prototipação.

S1 - Adorei o repeteco da Cultura "Especial Magda Tagliaferro".
S2 - O ex-VJ Thunderbird inaugura a nova programação teen da Globo.
S3 - Maiores informações podem ser obtidas na Tecnovídeo, representante da JVC no Brasil, pelo tel. (011) 816-6431.
S4 - Ligue grátis prá Claudette!
S5 - Pior, quando aquele brasileiro inpoliticamente incorreto viu o índio chegando, gritou: "Cuidado com o John Wayne".
S6 - E a Glória faz papel de cabra macho!
S7 - As siglas significam alta economia e alta performance, respectivamente.
S8 - Os outros Estados cresceram e em parte cresceram às custas de São Paulo.
S9 - Na estréia, um par de pentelhinhos argentinos dançando tango, Diego e Sabrina.
S10 - Tenho muita saudade dele e minha conta de telefônica fica obesa quando estamos longe.

Figura 2 – Corpus de sentenças escolhidas

Como podemos observar, temos sentenças simples, isto é, com somente um período, ou sentenças mais complexas, com mais de um período, as quais foram simplificadas e transformadas em sentenças simples (Figura 3). Essa foi outra limitação imposta ao nosso realizador, visando sua factibilidade. Após a etapa de simplificação, as sentenças foram analisadas sintaticamente pelo *parser* de Bick (2000)[2], que foi escolhido devido a seu alto índice de acerto nas construções sintáticas geradas.

SS1 - Adorei o repeteco da Cultura.
SS2 - Thunderbird inaugura a nova programação teen.
SS3 - Maiores informações podem ser obtidas na Tecnovídeo.
SS4 - Ligue grátis prá Claudette.
SS5 - Um brasileiro inpoliticamente incorreto viu o índio.
SS6 - A Glória faz papel de cabra macho.
SS7 - As siglas significam alta economia e alta performance.
SS8 - Cresceram às custas de São Paulo.
SS9 - Um par de pentelhinhos argentinos dançou tango.
SS10 - Minha conta telefônica fica obesa quando estamos longe.

Figura 3 – Corpus simplificado

Como resultado da análise, obtivemos as estruturas profundas, sintáticas, para dar origem às estruturas conceituais que servirão de entrada do protótipo. Adicionalmente, passamos a identificar as características de estilo do corpus simplificado, para definir os mapeamentos estilísticos fundamentais da proposta de realização superficial. Essas características são descritas a seguir.

4 Variações estilísticas em foco

A Tabela 1 mostra os itens lexicais de nosso corpus simplificado associados a variações estilísticas de naturezas diversas. Nessa tabela, estão descritas as figuras de estilo que o nosso protótipo contemplará, a saber: coloquialismo, corruptela, estrangeirismo, gíria, plebeísmo, adequação lexical, uso das locuções prepositivas e uso do comparativo. Essas figuras são descritas com detalhes em (Espina et al., 2002).

[2] Disponível para consulta em http://visl.hum.sdu.dk/visl/pt/.

Tabela 1 – Identificação das figuras de estilo

		Figuras de estilo							
Sen-tenças	Item(ns) lexical(is) em foco	Gíria	Estran-geirismo	Plebe-ísmo	Uso das locuções prepositivas	Adequação lexical	Colo-quia-lismo	Uso do compa-rativo	Cor-rupte-la
SS1	Repeteco	X							
SS2	Teen		X						
SS3	Maiores informações							X	
SS4	Pra						X		
SS5	Inpolitica-mente								X
SS6	Cabra macho			X					
SS7	Performance		X						
SS8	Às custas de				X				
SS9	Pentelhinhos			X					
SS10	Obesa					X			

A partir dessa tabela, definimos resumidamente cada um dos desvios estilísticos presentes no corpus:

1. Coloquialismo – Uso de vocabulário e sintaxe bem aproximados da linguagem do dia a dia.
2. Corruptela – Ato de escrever ou pronunciar erroneamente uma palavra ou locução.
3. Estrangeirismo - Uso de palavras ou expressões de outras línguas, como o galicismo ou anglicismo.
4. Gíria - Linguagem que, nascida num determinado grupo social, termina estendendo-se, por sua expressividade, à linguagem familiar de todas as camadas sociais.
5. Plebeísmo - Modos, usos, orações ou palavras de uso exclusivo da plebe.
6. Adequação lexical – Palavra que não está lexicalmente bem empregada.
7. Uso das locuções prepositivas – ‘Às custas de’ deve sempre ser trocado por ‘à custa de’.
8. Uso do comparativo – O uso do comparativo freqüentemente é empregado em contextos em que não se pretende comparar nada, como é o emprego de ‘maiores’em SS3. Neste caso, a forma correta é ‘mais informações’.

A partir dessa identificação dos problemas de estilo apresentados no corpus simplificado, pudemos identificar as modificações necessárias para transformá-las em sentenças aceitáveis, segundo as possíveis normas gramaticais da língua portuguesa.

Consideramos uma forma bastante simplificada para explorar tal transformação em nosso protótipo. Por exemplo, consideramos que, apesar de licenciadas (já que usadas nos próprios textos da Folha de São Paulo), as sentenças apresentadas corresponderiam, simplesmente, a uma norma menos culta do português, devendo, assim, ser mapeadas em uma norma mais culta. Essas variações de norma não seguiram o padrão clássico, isto é, foram determinadas intuitivamente neste trabalho. Por exemplo, na sentença SS6, na qual podemos identificar o plebeísmo correspondente à forma menos culta, deverá ser modificado para padrões mais aceitáveis que, no nosso entender, podem ser mais ou menos (MaMe) cultos ou mais (Ma) cultos (neste caso, essa seria, de fato, a norma culta nos padrões, por exemplo, do estado de São Paulo). A Tabela 2 apresenta esses possíveis mapeamentos de construções menos cultas em outras normas (Ma ou MaMe somente) a partir das figuras de estilo delineadas em nosso corpus.

Tabela 2 – Mapeamentos possíveis a partir da entrada

Estilo da Entrada	Norma Culta		Grau de Permissividade
	Ma	MaMe	Não permissivo
Coloquialismo	+	-	+
Corruptela	+	+	+
Estrangeirismo	-	-	+
Gíria	+	+	+
Plebeísmo	+	+	+
Adequação lexical	+	-	+
Uso das locuções prepositivas	+	+	+
Uso do comparativo	+	+	+

O grau de permissividade (3ª. coluna da tabela) corresponde à percepção de que não há uma rigidez extrema no uso da língua no Brasil. Assim, por exemplo, nosso protótipo terá, como fator de decisão adicional, a indicação do "comportamento" esperado na geração de uma nova forma sentencial: se o grau de permissividade for "+", a realização seguirá a norma mais culta esperada (Ma), mas com um fator adicional de não estar

permitindo nem mesmo a presença de estrangeirismos, o que é permitido pela norma culta. Se for "-", a realização seguirá a mesma norma de Ma.

Assim, essa tabela pode ser usada para definir os possíveis mapeamentos entre o estilo da entrada do realizador e suas supostas realizações alternativas. Por exemplo, para a gíria presente em SS1 (Figura 3), se quisermos a resolução 'Ma' (mais culta) e o grau de permissividade for '+', a sentença resultante será "Adorei a reprise da Cultura". Se, agora, a norma desejada for Ma (mais culta) e o grau de permissividade for '-', a sentença resultante será, neste caso, a mesma obtida anteriormente. Se agora a norma desejada for MaMe (mais ou menos culta) a sentença resultante será "Adorei a repetição da Cultura".

Para que o protótipo se comportasse do modo ilustrado, passamos à definição de seu repositório lexical e gramatical, conforme descrevemos nas seções seguintes.

5 Delimitação do Léxico

O léxico foi construído para representar todos os itens sugeridos no corpus simplificado (Figura 3). Utilizamos o léxico do NILC (Nunes et al., 1996) para importar as informações correspondentes, resultando no recurso 'Léxico do Português' indicado na Figura 1. Este recurso conta com 78 verbetes, no estágio atual do protótipo. Além de compreender os componentes sentenciais do corpus simplificado (somando 63 entradas do léxico), esse léxico tem mais 15 entradas, necessárias para o mapeamento apontado pelas regras de mapeamento estilístico (vide Figura 1).

A marcação do léxico do NILC foi alterada para atender as necessidades do nosso realizador, pois, além de não ter informações estilísticas, continha várias informações desnecessárias, como, por exemplo, regras de derivação de gênero e número, as quais não são utilizadas atualmente em nosso protótipo. Incluímos, assim, as informações relativas a estilo, mas também excluímos as informações que eram desnecessárias. As Figuras 4 e 5 ilustram, respectivamente, a forma original das entradas do léxico e a forma proposta em nosso protótipo, de forma genérica, em função das categorias léxicas do português.

1. Substantivos ➔ subs(Gênero, Número, Grau, Regência do Substantivo, Regra Deriv. Gênero, Regra Deriv. Número, Forma Canônica)
2. Adjetivo ➔ adj(Gênero, Número, Grau, Regência do Adjetivo, Regra Deriv. Gênero, Regra Deriv. Número, Forma Canônica)
3. Artigo ➔ art(Gênero, Número, Tipo, Regra Deriv. Gênero, Regra Deriv. Número, Forma Canônica)
4. Preposição ➔ prep(Contração, Forma Canônica)
5. Conjunção ➔ conj(Tipo, Complemento Coordenativa, Complemento Subordinativa, Forma Canônica)
6. Pronome ➔ pron(Gênero, Número, Tipo, Pessoa, Regra Deriv. Gênero, Regra Deriv. Número, Contração, Forma Canônica)
7. Verbo ➔ v(Predicação do Verbo, Formas Nominais, Par(Tempo, Pessoa) Colocação Pronominal, Regência do Verbo, Forma Canônica)
8. Advérbio ➔ adv(Tipo, Grau, Forma Canônica)

Figura 4 – Traços morfossintáticos originais

1. Substantivos ➔ subs(Número,Gênero,Norma de Estilo,Forma Canônica)
2. Adjetivo ➔ adj(Número,Gênero,Norma de Estilo,Forma Canônica)
3. Artigo ➔ art(Tipo,Número,Gênero,Forma Canônica)
4. Preposição ➔ prep(Norma de Estilo,Preposição Canônica,Artigo Canônico)
5. Conjunção ➔ conj
6. Pronome ➔ pron(Número,Gênero)
7. Verbo ➔ v(Número,Pessoa,Tempo,Forma Canônica)
8. Advérbio ➔ adv(Norma de Estilo)

Figura 5 – Traços morfossintáticos do protótipo

As Figuras 6 e 7 mostram a representação para itens lexicais específicos do corpus simplificado.

1. Substantivo ➔ programação=<S.F.SI.N.[]?.1.[programação]0.>
2. Adjetivo ➔ nova=<ADJ.F.SI.N.[de.]1.3.[novo]0.>
3. Artigo ➔ a=<ART.F.SI.DE.?.?.[o]0.#PREP.[a]0.#PRON.F.SI.[DEM. OBL-AT.]3S.?.?.C.[][o]0.#ABREV.M.SI.[a]0.#S.M.SI.N.[]?.?.[a]0.>
4. Preposição ➔ às=<PREP.C.[a.as.][ao]0.>
5. Conjunção ➔ e=<CONJ.[COORD.[ADIT.ADVE.]][e]0.>
6. Pronome ➔ minha=<PRON.F.SI.[POSS.]1S.?.3.C.[][meu]0.>
7. Verbo ➔ inaugura=<V.[][IMPER-AFIRM.TU.PRES.ELE.]N.[] [inaugurar]0.>
8. Advérbio ➔ longe=<ADV.[CIR-LUG.]N.[longe]3.#ADJ.2G.SI.N.[]? .3.[longe]3.>

Figura 6 – Entradas lexicais originais

1. Substantivo ➔ programação=<sing,fem,n,programação>
2. Adjetivo ➔ nova=<sing,fem,n,nova>
3. Artigo ➔ a=<def,sing,fem,a>
4. Preposição ➔ às=<n,a,as>
5. Conjunção ➔ e=<>
6. Pronome ➔ minha=<sing,fem>
7. Verbo ➔ inaugura=<sing,terc,pres,inaugurar>
8. Advérbio ➔ longe=<n>

Figura 7 – Entradas lexicais do protótipo

6 Delimitação da gramática

Como consideramos somente as construções simples como as ilustradas na Figura 3, nossa sub-gramática pode ser expressa, de forma genérica, pelas regras gramaticais apresentadas na Figura 8[3].

```
ac --> adv,v,adv.
pp --> prep,subs.
pp --> prep,sns.
s --> sns,sv.
s --> sns,sv,pp.
s --> sv.
sns --> adj,subs.
sns --> art,adj,subs,adj.
sns --> art,subs.
sns --> art,subs,adv,adj.
sns --> art,subs,pp.
sns --> subs.
sns --> subs,adj.
sns --> subs,pp.
sns --> pron,subs,adj.
snc --> sns,conj,sns.
sv --> v,adj,ac.
sv --> v,adj,pp.
sv --> v,sns.
sv --> v,snc.
sv --> vaux,v.
vaux --> v,v.
```

Figura 8 – Exemplo de Sub-gramática

7 Considerações finais

Como vimos, a partir da identificação das regras de estilo que deveriam ser contempladas, traçam todos os recursos lingüísticos necessários para construir nosso protótipo. Neste caso, temos a seguinte configuração (vide Figura 1):

- As regras de estilo servirão para identificar, na estrutura conceitual de entrada, qual o problema de estilo em questão (recurso construído com base na Tabela 1);
- As regras de mapeamento servirão para identificar as possíveis variações da entrada (recurso construído a partir da Tabela 2);

[3] ac → cláusula adverbial; pp → sintagma nominal preposicionado; s → sentença; snc → sintagma nominal composto; sns → sintagma nominal simples; sv → sintagma verbal; vaux → verbos auxiliares

- O léxico, assim como a sub-gramática, completam o conjunto dos quatro recursos lingüísticos do protótipo.

A etapa seguinte à especificação lingüística do realizador foi baseada na construção de todas as estruturas-f das sentenças do corpus escolhido para, então, detalhar as regras de mapeamento estilístico em função dessa linguagem interna de representação conceitual.

No estágio atual do sistema, consideramos somente variações lexicais, como ilustramos nesse relatório. Isto significa dizer que as estruturas-f de entrada do realizador serão as mesmas estruturas-f das sentenças a serem produzidas, porém, estas irão contemplar variações lexicais. Equivalentemente, isto é o mesmo que manter a estrutura sintática da entrada.

Embora essa proposta seja simples, é possível, como já mencionamos antes, aumentar o poder lingüístico-computacional do realizador de diversas formas. Por exemplo, podemos aumentar o número de sentenças de entrada ainda mantendo a mesma sub-gramática. Neste caso, o protótipo não será mais expressivo em termos sintáticos, mas possibilitará a geração de outras sentenças. Estas, por sua vez, podem estar associadas a outros casos de estilo, situação em que os próprios conjuntos de regras de estilo e de mapeamento deverão ser estendidos.

Para essas duas situações, considerar novos dados de entrada não parece implicar um esforço muito grande. Entretanto, a inclusão de novos dados que impliquem alterações sintáticas é mais complicada, pelo custo de estender a sub-gramática e as regras de mapeamento estilístico, as quais devem compreender, agora, as descrições das variações sintáticas também.

Referências Bibliográficas

Bick, E. (2000). *The parsing system Palavras: automatic grammatical analysis of Portuguese in a constraint grammar framework.* Arhus University Press. Tese de Doutorado.

Espina, A.P.; de Bem, M.J.C.; Rino, L.H.M. (2002). *A exploração de questões de estilo do português para a realização superficial automática.* Série de Relatórios Técnicos do NILC, NILC-TR-02-16. Setembro.

Kaplan, R. & Bresnan, J. (1982). Lexical-Functional Grammar: a Formal System for Grammatical Representation. In Bresnan, J. (ed.), *The Mental Representation of Grammatical Relations*, pp. 173-281. MIT Press, Cambridge MA.

Kuhn, D., Abarca, E., Nunes, M.G.V.(2000). *Corpus NILC - Situação em Maio/2000.* Série de Relatórios Técnicos do NILC, NILC-TR-00-7. Junho, 32p.

Nunes, M.G.V. et al. (1996). *A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos*. Relatórios Técnicos do ICMC-USP, 42. Setembro, 36p.