

Processamento Linguístico Aplicado à Síntese da Fala

Daniela Braga*, Diamantino Freitas, Helder Ferreira

dbraga@ese.ipp.pt, dfreitas@fe.up.pt, hfilipe@fe.up.pt

*Escola Superior de Educação – Instituto Politécnico do Porto, Faculdade de Engenharia da Universidade do Porto

Resumo: Neste trabalho pretendemos mostrar como o conhecimento fornecido pela linguística teórica pode contribuir para o sucesso dos sistemas de síntese da fala, utilizando uma linha de programação de algoritmos segundo regras. Entre as principais linhas de intervenção da linguística aplicada à síntese da fala, constam o pré-processamento do texto escrito (com tarefas de transformação, como a conversão de grafemas, numerais, siglas e acrónimos) e o processamento prosódico subsequente, através da relação entre os níveis morfo-sintáctico e prosódico da frase. Pretende-se igualmente apresentar um exemplo de integração do conhecimento linguístico num sistema de Text-to-Speech, construído no contexto da experiência desenvolvida no Laboratório de Processamento da Fala da FEUP (<http://lpf-esi.fe.up.pt>).

1. INTRODUÇÃO

A Linguística tem como objecto de estudo o uso/ funcionamento das línguas e da linguagem humana. Em linguística procura-se descrever o conhecimento intuitivo e inconsciente que permite que um falante use uma língua para comunicar com o outro. Uma língua é o resultado da combinação do Léxico com a Gramática, ou seja do vocabulário com regras de organização das unidades que constituem a língua.

Para falar uma língua, é necessário conhecer o conjunto de elementos sonoros que a compõem, saber como combiná-los e que sons não podem ser agrupados. O conhecimento que permite decidir que a sequência fónica [‘patu] é possível em português, mas que a sequência [‘fst] não pertence ao sistema sonoro do português, constitui o conhecimento fonético de uma língua. De igual modo resultam deste nível de conhecimento a formulação de regras de restrição e de dependência motivadas por vizinhanças contextuais, como a diferença entre a articulação uvular múltipla de um <r> em início de palavra, por exemplo em [‘Ratu], e a articulação dental simples de um <r> em final de sílaba, como em [‘mar].

A Fonologia estuda os aspectos sonoros que distinguem significados numa língua. Para a fonologia, não existe diferenciação entre uma articulação dental ou velarizada do grafema <l> em português, fenómeno que, por exemplo, tem que ser considerado em fonética, pois existe uma obrigatoriedade contextual e fisiológica que nos leva a articular o grafema <l> em <mal> e em <lobo> de maneiras diferentes. A Fonologia desenvolveu uma linguagem formal para descrever a realidade fonética significativa da língua.

As palavras são compostas por unidades menores, os morfemas. Por exemplo, na palavra <folhagens>, sabemos que existe uma palavra-raiz, a palavra *folha*, de onde ocorreu uma derivação pela junção de um sufixo *-agem*, com um valor de colectivo. Além disso, o sufixo final <-s> adiciona uma informação número, neste caso de plural. Informações sobre os processos de formação de palavras inserem-se no conhecimento morfológico de uma língua.

Considerem-se agora as seguintes frases:

- i. *A maior parte das pessoas compram o JN.
- ii. *Tenho pena que isto é assim.
- iii. *Nota-se grandes diferenças nesta cidade.

Qualquer falante escolarizado consideraria estas frases agramaticais, ou seja, com algum tipo de má formação. O nível de conhecimento aqui implicado é o nível sintáctico, que é o que permite reconhecer em i) uma não concordância entre o sujeito singular e o predicado (que está no plural por proximidade de um antecedente plural que não é o sujeito). De igual modo é o conhecimento sintáctico que explica a agramaticalidade de ii), uma vez que o verbo da oração subordinada ‘é’ devia estar não no Indicativo, mas no modo Conjuntivo devido às restrições impostas pela estrutura ‘Tenho pena’. Em iii), a inversão do sujeito ‘grandes diferenças’ colocou-o na posição normal ocupada pelo complemento directo, pelo que não se verificou a necessidade de fazer a concordância entre o sujeito e o predicado, como acontece sempre em português.

A Semântica é o estudo do significado de uma língua. É o que permite responder a uma questão num exame ou entender um livro de instruções. Permite-nos perceber relações associativas entre ‘viagem’, ‘aeroporto’, ‘avião’, ‘praias’, e saber que a palavra <banco> pode significar, de acordo com o contexto, quer ‘*instituição bancária*’, quer ‘*objecto que serve de assento*’. Problemas de homonímia ou polissemia como este último constituem dificuldades de grande importância para o tratamento computacional do texto.

Na conversação quotidiana, convocamos todo o nosso conhecimento linguístico e usamo-lo para comunicar. Contudo, esse conhecimento não é suficiente face à complexidade envolvida na comunicação interpessoal. Precisamos de ter em conta a idade, o estatuto socio-económico do nosso interlocutor, o tipo de relação que temos com ele. Evocamos tudo o que sabemos sobre o nosso interlocutor, os pressupostos, as crenças, o universo de referências comum a ambos. Este conjunto de conhecimentos constitui o que se entende por Contexto, e está na base da interpretação de um enunciado do tipo ‘*Podia dizer-me as horas?*’, não como uma frase interrogativa, de onde se espera uma resposta sim/ não, mas sim um pedido, de onde se espera a obtenção de uma informação.

Ora, uma vez que os sistemas de CTF (conversão texto-fala, do inglês: text-to-speech ou só TTS) partem do suporte escrito de uma língua com o propósito de gerar a sua componente oral, surgiu a necessidade de estreitar a relação entre o conhecimento sobre as línguas fornecido pela linguística e o seu processamento e integração num sistema informático pela engenharia. Os capítulos seguintes apresentam as etapas que em se desenvolveu esse processamento linguístico aplicado.

2. CONSTITUIÇÃO DA BASE DE DADOS

Subjacente à arquitectura de alguns sistemas de síntese da fala, como o de concatenação de unidades, tem que existir uma base de dados constituída por gravações de voz natural que, depois de etiquetada, fornecerá os difones ou outros segmentos maiores, que combinados,

constituirão a voz sintética. Os critérios que subjazem à construção e etiquetagem de uma BD estão dependentes da técnica de síntese que se pretende utilizar e da qualidade que se pretende atingir. De qualquer modo, a linguística teórica pode fornecer os instrumentos de análise de forma a conseguir-se rentabilizar melhor o trabalho de preparação da base de dados.

Em primeiro lugar, devem seleccionar-se textos (lidos)/ diálogos que sejam suficientemente representativos da diversidade e riqueza fonética, sintáctica e prosódica da língua. Deve também ser tida em consideração a qualidade da gravação, que deve ser realizada em condições especiais com vista a obter-se a maior qualidade possível na síntese. A escolha do informante (Locutor) é também fundamental, pois deve ter boa articulação, timbre de voz agradável e falar uma variante fonética próxima da norma/ variante de prestígio, excepto se o objectivo do projecto de síntese for o de produzir sinteticamente uma variante dialectal que não coincida com a normativa.

Após a recolha dos sons da futura base de dados, é necessário proceder-se à selecção dos elementos e factores linguísticos que devem ser assinalados e marcados para poderem servir de plataforma de trabalho para a programação de regras no sistema CTF, ou seja, é preciso proceder-se à etiquetagem do sinal de fala. Normalmente marcam-se fronteiras entre os fonemas, bem como as pausas, e assinalam-se os elementos prosódicos, em especial as durações e as oscilações da frequência fundamental, f_0 . Esta etiquetagem mostrou-se ser uma tarefa árdua, apesar da relativa eficácia dos recursos informáticos disponíveis para este efeito, uma vez que não há unanimidade de critérios a seguir na definição de categorias, como por exemplo em relação ao momento onde se marca uma fronteira de fonema, ou em relação aos elementos e factores prosódicos a considerar. Verifica-se, portanto, a ausência de uma simbologia consensual que descreva com rigor todos os fenómenos, e em particular, os fenómenos suprasegmentais do domínio da entoação, do foco, do ritmo e do estilo individual.

Existem alguns laboratórios que desenvolveram bases de dados com etiquetagem linguística (LPL – Laboratoire de Parole et Langage (www.lpl.univ-aix.fr); LAIP – Laboratoire pour la synthèse de la parole (www.unil.ch/imm/docs/LAIP/LAIP_TTS_fr.htm); ACULAB (www.aculab.com).

A nossa experiência produziu também uma base de dados constituída por 20 minutos de voz etiquetada ao fone, sílaba, palavra e frase, de que um extracto pode ser encontrado em www.portugues.mct.pt/Repositorio/EuroSpeechIPB/.

3. PRÉ-PROCESSAMENTO DO TEXTO

O pré-processamento de texto envolve todas as tarefas que estão na base da descodificação do texto. Um sistema de Conversão Texto-Fala parte de um texto escrito que deverá interpretar e converter em voz, à semelhança do procedimento psico-cognitivo-articulatório que é desenvolvido pelo ser humano.

3.1. Conversão grafema-fonema

As línguas são constituídas por três tipos de sons, de acordo com as características da sua produção: consoantes, vogais e semi-vogais. Cada língua selecciona uma lista fechada de sons que combina segundo regras e de acordo com certos contextos.

A invenção da escrita teve como principal objectivo fornecer um suporte visual e permanente à língua oral. Apesar da escrita da língua portuguesa ter sido originariamente de natureza alfabética, isto é, baseada numa relação unívoca entre um som e um grafema, a verdade é que essa relação não se verifica actualmente, devido à tensão existente entre o dinamismo da evolução fonética da língua e o carácter conservador, próprio de todos os sistemas de escrita.

Surgem por consequência, dificuldades de conversão de alguns grafemas, e em especial dentro do vocalismo do português, como veremos em seguida. Essa dificuldade estende-se à transcrição das palavras homógrafas, como <sede> [‘sedĩ] e <sede> [‘sedĩ], cuja resolução depende da programação de regras de desambiguação através do recurso a um dicionário.

Outra questão subjacente à conversão grafema-fonema diz respeito à selecção do sistema de transcrição, que esse sim, deve ser unívoco. Com vista à sua implementação por computador, tem-se adoptado o sistema SAMPA de representação fonética, que apesar de menos completo do que o sistema IPA, é o mais adequado a este tipo de aplicação por se servir dos caracteres disponíveis no teclado do computador (PC).

A transcrição pode realizar-se de várias formas, entre as quais se citam, por regras e por classificadores/conversores estatísticos (redes neuronais artificiais, por exemplo). No LPF-ESI exercitam-se ambos os tipos de abordagem, estando todavia mais desenvolvido o primeiro método. Apresentamos, em seguida, as regras de conversão principais para o PE.

3.1.1. Regras de conversão de consoantes

Existem 19 fonemas consonânticos em português (Mateus, 1975) e 18 consoantes gráficas. Isto prova que não é possível uma transcrição directa de letras para sons, além de que existem, por exemplo, grafemas que não têm correspondência sonora (como o <h>, que perdeu a aspiração durante a evolução da língua, como em <hora>), grafemas que convergem para o mesmo som (como o <s> e <c> que têm valor de [s], como em <silo> e em <ciclo>) ou como o <g> e o <j> que se realizam como [ʒ] em <geral> e <jeito>) e grafemas que têm duas articulações diferentes consoante os contextos fonológicos (<s> que se realiza como [z] em posição intervocálica (<casa>), como [ʒ] antes de consoante sonora (<osga>) e como [ʃ] antes de consoante surda (<casca>).

Foi possível elaborar um conjunto de regras de conversão grafema-fonema suficientemente eficaz para o consonantismo do português no decorrer da nossa experiência de construção de sistemas de síntese no LPF-ESI. Apenas o grafema <x> se mostrou mais irredutível a regras em virtude da sua evolução histórica (o grafema <x> pode realizar-se de quatro modos: [ks] em <tórax>, [s] em <próximo>, [z] em <exame> e [ʃ] em <xaile>), pelo que apesar de terem sido encontradas algumas regularidades de conversão, muitas palavras apresentam transcrições fonéticas imprevisíveis, o que nos levou a criar uma tabela de excepções para este grafema.

3.1.2. Regras de conversão de vogais

A não verificação de uma relação unívoca entre grafemas e fonemas repete-se no que se refere às vogais do português, sobretudo quando se considera o vocalismo oral. Na verdade, existem 9 vogais orais representados por apenas 5 grafemas vocálicos. O grafema <e> é o que apresenta maior espectro de realizações fonéticas, com pelo menos 5 possibilidades: <sebe> → [ɛ], <selo> → [e], <secar> → [i], <seja> → [a] e <estado> → [ø]. Além disso, as semi-vogais [j] e [w] são também representadas pelos grafemas <i> e <u>, o que contribui para aumentar a ambiguidade gráfica. Assim, seguiu-se a solução adoptada para o grafema <x>, ou seja, criou-se uma tabela de excepções, para palavras que escapam às regras, sempre aberta a novas inclusões.

O vocalismo nasal apresenta, apesar de tudo, um sistema de transcrição mais unívoco, ainda que tenha que servir-se de dígrafos para a sua representação (<on->/ <om-> → [õ], como por exemplo em <onde> e <sombra>).

3.1.3. Regras de coarticulação

Durante a produção vocal, nunca articulamos o mesmo som rigorosamente da mesma forma. No entanto, as principais características distintivas mantêm-se, assegurando a inteligibilidade na língua. Existem, porém, vizinhanças fonéticas que condicionam e introduzem restrições de ordem articulatória, produzindo diferenças acústicas significativas, embora sem consequências no plano do significado das palavras, mas que devem ser assinaladas pelos sistemas de síntese, sob pena de provocar prejuízos na naturalidade da fala sintetizada. Deste modo, é importante proceder-se ao levantamento dos fenómenos de coarticulação, encontrar neles regularidades e reconhecer eventuais variações alofônicas de alguns fonemas, para em seguida os etiquetar e os adicionar à lista de difones da base de dados. Seguem-se alguns exemplos de fenómenos de coarticulação em Português europeu (PE) que dão origem ao aparecimento de variações contextuais de alguns fonemas:

- 1) A consoante lateral dental /l/, quando ocorre em final de sílaba sofre velarização em PE [ɫ]:
 - i. <sal> → [ˈsa ɫ], <salgado> [saɫˈɡadu]:
- 2) As consoantes oclusivas sonoras /b/, /d/, /g/ em posição intervocálica fricativizam [β], [ð], [ɣ]:
 - i. <saber> → [s α ˈβer], <Sado> → [ˈsa ð u], <saga> → [ˈsa γ α]
- 3) A consoante oclusiva dental surda /t/ em sílaba aberta final de palavra antes de <e> átono fricativiza [ts], acompanhada de apócope da vogal átona final:
 - i. <noite> → [ˈnojts]

Além de fenómenos relacionados com consoantes, ocorrem também inúmeros processos de supressão de vogais átonas em PE. Estudos recentes em fonética têm demonstrado que existe uma mudança fonética em curso, traduzida por uma tendência para o enfraquecimento do vocalismo átono do português da zona de Lisboa (variante normativa), enfraquecimento esse que surge em «*vários graus, que vão de um simples relaxamento até à queda dos sons, passando por processos claros de redução fonética*» (Delgado-Martins, 2002: 302). Contudo, não existem ainda estudos suficientes que reduzam estes fenómenos a regras, com exceção dos ocorridos em alguns contextos silábicos (antes de consoante fricativa [ʃ] – cfr. Andrade, 1994; redução do /e/), até porque uma vez que se trata de uma mudança ainda em curso, há a coexistência das duas realizações sem prejuízo de nenhuma. A queda de vogais provoca encontros consonânticos anteriormente não permitidos na língua portuguesa, o que tem como consequência uma total reestruturação silábica.

Frota (2000) aborda também, no seu trabalho sobre o foco prosódico, fenómenos de *sandhi* em PE, que não são mais do que fenómenos de coarticulação. Eis alguns dos fenómenos detectados:

- 1) Sonorização da fricativa [ʃ]
 - a. em sílaba fechada, antes de consoante sonora realiza-se como [ʒ]
 - i. musgo [ˈmuʒgu]
 - b. em sílaba fechada, antes de vogal realiza-se como [z]
 - i. lápis azuis [ˈlapiz α zuj ʃ]

- 2) Redução vocálica e degeminação consonântica
 - i. campo **p**equeno → cam[p p]equeno
- 3) Crase (Fusão vocálica)
 - i. **A** **a**luna **a**ceitou o emprego. [α α] → [a]
- 4) Redução da vogal final
 - i. esta imagem [αi] → [i]
- 5) Semivocalização
 - i. salto **a**lto [u a] → [wa].

3.2. Conversão de numerais, siglas, acrónimos e abreviaturas

Um módulo de síntese de voz não distingue as palavras de números, não sabe o que são abreviações, não sabe o que são acrónimos, não entende semântica. Sendo assim, para um sintetizador, uma frase não passa de um conjunto de caracteres sem significado. É então necessário introduzir marcações ou regras, que permitam identificar e desambiguar determinadas situações, para que o sistema seja capaz de ler correctamente a palavra ou até mesmo a frase. Por exemplo, tomemos a seguinte frase: “*O prof. João dá aulas, às 9h00 na sala 123, na FEUP.*”. Esta frase contém uma sigla (“*FEUP*”), uma abreviação (“*prof.*”), e dois tipos de representação numeral: um tempo (“*9h00*”) e um cardinal (“*123*”). Todos estes elementos têm de ser convertidos para a sua forma em extenso de modo a que um conversor texto-fala possa “ler” a frase correctamente. Colocam-se agora dois problemas distintos: como identificar estes elementos numa frase e como convertê-los?

3.2.1. Identificação de elementos

Existem duas formas de identificar os elementos anteriormente referidos: usando uma técnica de etiquetagem prévia ou fazendo reconhecimento automático. No entanto, tanto uma como outra exigem um conhecimento anterior do tipo de elementos que podem surgir numa frase e de uma classificação dos mesmos, para que se possam estabelecer regras, proceder a marcações ou procurar padrões.

3.2.1.1. Tipos e classificações

Existem diversos tipos de elementos num texto que necessitam de conversão, sendo os mais comuns: os numerais, as abreviações, as siglas e os acrónimos. Em documentos de carácter científico podemos ainda encontrar diversos elementos e símbolos matemáticos, para além dos anteriormente referidos. Actualmente começam a surgir novos tipos de elementos, tais como: endereços de sites ou endereços de correio electrónico. Veja-se na Tabela 1 uma classificação dos tipos mais comuns.

Tabela 1 – Tipos de elementos de interesse para o pré- processamento do texto.

Classe	Sub-classe
Numerais	cardinais, ordinais, decimais, fraccionários, romanos, telefones, datas, horas, resultados de jogos, apostas, escalas, potências, dízimas, complexos, indexações, listagens, códigos, monetários, outros.
Siglas e Acrónimos	organizações, projectos, outros.

Abreviações	unidades, social, outros.
Referências Web	url, uri, ftp, email, outros.

3.2.1.2. Etiquetagem

Depois de conhecermos os principais tipos de elementos que existem e quais as suas classificações, podemos agora desenvolver um conjunto de etiquetas que permitam marcar o elemento em questão agregando-lhe um conjunto de informação relevante para a conversão e para o desambiguação de determinadas situações. Por exemplo, consideremos a seguinte frase: “*Falei com o João. Foi 2:1.*”. Nesta frase não se conhece o contexto, pelo que existem duas possibilidades de leitura distintas: “...Foi dois um.”, no caso de ter sido um resultado de um evento desportivo, p.ex. futebol; “Foi dois para um.”, no caso de se referir a uma escala de um mapa ou esquema. No caso de ausência de etiquetagem tem de existir um modo de leitura por omissão. Existindo etiquetas e marcações, por exemplo, ‘*<numeral tipo=”escala”>2:1</numeral>*’, poderíamos “ler” correctamente a frase.

Actualmente existem diversas linguagens de marcação disponíveis, sendo as mais importantes: SSML (*Speech Synthesis Markup Language*), SABLE, VoiceXML (*Voice Extensible Markup Language*) e MathML (*Mathematical Markup Language*). No entanto, estas linguagens ainda se encontram em desenvolvimento e ainda não reconhecem todas as sub-classes apresentadas no ponto anterior. Sendo assim, o LPF-ESI desenvolveu uma linguagem de marcação para o pré-processamento de textos num conversor texto-fala, denominada: TPML (*Text Processing Markup Language*).

3.2.1.3. Reconhecimento automático

Outra forma de identificar os elementos num texto, é através da procura de padrões usando máquinas computacionais de estados finitos (autómatos de estados finitos ou transdutores de estados finitos) programadas em C ou noutra linguagem, ou simples máscaras de regras que filtram o texto. Por exemplo, se se encontrar um algarismo delimitado por espaços, está-se na presença de um número do tipo cardinal. Se se encontrar um espaço seguido de um conjunto de algarismos, seguido do símbolo “o” ou do símbolo “a”, então está-se na presença de um número do tipo ordinal.

A segunda técnica é a mais simples, e proporciona bons resultados para os tipos de sub-classes mais comuns (cardinais, ordinais, monetários, percentagens). No entanto, torna-se demasiado complexa para alguns tipos como por exemplo, identificar um número romano no meio de uma frase. Para estes casos, o uso de máquinas de estados finitos é o mais indicado.

3.2.2. Conversão dos elementos

Para realizar a conversão dos elementos (numerais, abreviações, acrónimos...) necessitamos de elaborar diversos mecanismos que sejam aplicados após a identificação da classe e sub-classe do elemento. Estes mecanismos consistem num conjunto de funções ou algoritmos cuja entrada seja o elemento em causa, e a saída uma representação por extenso do mesmo.

Geralmente estes algoritmos baseiam-se na concatenação de unidades de texto que seguem determinadas regras da gramática. Por exemplo, suponhamos que na entrada do algoritmo temos o número “22”. Existe uma regra que diz que se o número é cardinal e é composto por dois algarismos, e se o algarismo da esquerda é um 2, então dizemos “vinte e” e fazemos concatenar com a representação por extenso do algarismo da direita que neste caso é “dois”. O resultado seria “vinte e dois”.

É comum desenvolverem-se tabelas que relacionam a representação simbólica com a representação por extenso, e depois aplicarem-se simples regras como a que foi demonstrada anteriormente.

3.3. Regras de divisão silábica

A sílaba é uma unidade relativamente fácil de identificar e de segmentar se seguirmos as regras de divisão silábica estipuladas pela ortografia portuguesa. Essas regras foram já implementadas por Teixeira e Freitas (2000) num sintetizador de voz.

Contudo, enquanto unidade psicológica e fonológica, é constituída por um agrupamento fónico em que a vogal constitui o pico e o núcleo da sílaba, por ser o som que apresenta o grau mais elevado de sonoridade (M. João Freitas, 2002). Na vizinhança esquerda e direita da sílaba ocorre um inventário de consoantes assimétrico, uma vez que está sujeito a restrições de distribuição, ou seja, a consoante lateral velarizada [ɫ] não pode ocorrer em ataque silábico, podendo apenas surgir em coda (ex: <mal>). A sílaba recebe ainda um determinado acento de intensidade, com uma certa duração enquanto o seu contorno entoacional depende do tipo de frase. A estrutura da sílaba em PE, estudada por M. J. Freitas e Mateus, encontra-se, no entanto em revisão, uma vez que na sequência dos fenómenos de queda do vocalismo átono, a sílaba tem sofrido uma reestruturação profunda, devido aos novos encontros consonânticos que daí resultaram, o que tem equacionado a atribuição de um novo estatuto às consoantes, que à semelhança de outras línguas, passariam a ser núcleos de sílabas (Delgado-Martins, 2002: 280).

A importância da definição de regras de divisão silábica e de identificação das sílabas tónicas para um sistema de conversão Texto-Fala decorre da necessidade da identificação da alternância entre sílabas tónicas e átonas nas palavras, processo que marca o ritmo das frases e que afecta o nível prosódico, na medida em que uma sílaba tónica apresenta um aumento de intensidade e de duração. Além disso, a marcação da sílaba tónica pode ser responsável pela desambiguação de palavras como <explícito> e <explicito>, <fábrica> e <fabrica>, <cópia> e <copia>, processo que tem inclusive repercussões morfosintácticas como a alteração da categoria da palavra [<explícito> (nome ou adjectivo)/ <explicito> (verbo)].

4. PROCESSAMENTO PROSÓDICO

A conversão da ortografia do texto para som constitui uma parte do *input* de um sistema de síntese. Outra componente importante para a geração de voz é a prosódia. A prosódia é o nível fonético da língua que não é segmentável em unidades discretas, como o são os fones, apesar de ser o aspecto linguístico que confere naturalidade à fala e que realmente distingue o homem da máquina. A prosódia engloba as características de entoação, de intensidade e de duração fonética, e em português está na base de distinções fonológicas importantes, como a alternância entre sílabas tónicas e átonas, ou como a diferença entre frases declarativas, interrogativas ou exclamativas. É também responsável pela distinção perceptiva entre informação nova e importante e informação conhecida e menos relevante num texto/discurso. Exprime a atitude do falante perante o conteúdo proposicional do enunciado. Reflecte ainda toda uma semântica das emoções.

4.1. Análise linguística do texto

A análise linguística representa o segundo nível de processamento com vista à manipulação prosódica em síntese da fala. Em seguida apresentam-se alguns procedimentos gerais que estão na base desta etapa de processamento do texto.

4.2.1. Tratamento da pontuação

A pontuação não é senão um sistema rudimentar de transcrição prosódica, fornecendo as primeiras fronteiras dentro das frases e entre frases. No entanto, acontece frequentemente que a pontuação não dá conta de todos os fenómenos prosódicos, como por exemplo, o fenómeno da pausa entre o sujeito e o predicado, nem tão pouco de contornos entoacionais distintos transcritos com o mesmo sinal de pontuação, como a curva descendente de uma frase interrogativa Qu- (*‘Como conseguiste a bolsa?’*) e a curva ascendente de uma frase interrogativa Sim/Não (*‘Conseguiste a bolsa?’*). Apresentando a pontuação este tipo de limitações, surgiu a necessidade de fundamentar o processamento prosódico em algo mais sistemático e rigoroso, que não se ancorasse apenas no nível gráfico. Desenvolveu-se, portanto, um modelo de análise linguística do texto a partir da classificação das categorias morfossintáticas das palavras e da sua organização em unidades sintáticas e semânticas, de onde finalmente decorrem os modelos prosódicos.

4.2.2. Do plano morfossintático ao prosódico

As abordagens linguísticas do processamento da fala partem de uma explicitação de regras que se baseiam no pressuposto da existência de uma correlação entre o conhecimento morfossintático do texto e a modulação prosódica da voz (Monaghan, 2000). Por outras palavras, os contornos entoacionais, a intensidade e duração que imprimimos sobre alguns segmentos verbais decorrem da natureza e tipo de expressões/frases que articulamos, ou seja, dos níveis morfológico, sintático e semântico da língua. E como o verbo em português é a categoria central da frase, já que, enquanto predicador, determina as funções sintáticas e semânticas dos argumentos, o seu número, estrutura e distribuição, desenvolvemos um modelo de análise linguística inspirado nos princípios da Gramática de Valências (Vilela, 1999) que parte do verbo para a análise da frase e dos seus constituintes.

Esta análise gramatical necessitou de recursos informáticos para a identificação e classificação morfossintática das palavras, pelo que se recorreu a um analisador morfológico cedido por uma empresa de ferramentas linguísticas de suporte informático - PRIBERAM.

Uma vez reconhecidos os verbos da frase e identificadas as categorias das restantes palavras, a gramática de regras sintáticas vai encontrar constituintes e atribuir-lhes as possíveis fronteiras sintáticas. Esta gramática foi programada em PROLOG, mas pode sê-lo também em C ou noutras linguagens.

O passo seguinte é a organização desta informação linguística em grupos frásicos, ou seja, unidades sintático-semânticas maiores que os sintagmas anteriores. Um grupo frásico é uma unidade de sentido, delimitada por uma fronteira sintática a que corresponde um determinado modelo entoacional normalmente separado por uma pausa. Por conseguinte, a cada grupo frásico corresponderá um grupo prosódico com um certo padrão de f0 de acordo com a posição e tipo de frase em que se insere.

Para ilustrar este procedimento, consideremos a seguinte frase:

O marquês e Taveira moviam lentamente as pedras, sem uma palavra, com um ar de rancor surdo. (apud Eça de Queirós, *Os Maias*).

Nesta frase, a pontuação é um factor importante de demarcação de fronteiras entre grupos frásicos, mas não prevê, por exemplo, a pausa, ainda que facultativa, entre o sujeito composto *‘O marquês e Taveira’* e o predicado verbal *‘moviam lentamente as pedras’*. Após ter sido identificado o verbo e terem-se agrupado as palavras da sua vizinhança em sintagmas, a nossa gramática, de acordo com a tipologia de que dispõe até ao momento, apresentaria a análise que pode ver-se na Tabela 2:

O inventário de regras que articula os grupos frásicos com os padrões prosódicos decorre da análise do contexto distribucional das categorias das palavras na gramática. A cada grupo frásico faz-se corresponder um grupo prosódico. Os padrões prosódicos diferem de acordo com a sua posição não final ou final na frase. Todos os grupos frásicos não finais apresentam contornos semelhantes, independentemente do tipo de frase em que ocorrem. É justamente ao nível do grupo frásico final que se verificam padrões entoacionais pragmaticamente distintivos (declarativo, declarativo-suspensivo, interrogativo QU-, interrogativo sim/não, interrogativo tag, interrogativo eco, interrogativo alternativo, exclamativo de emoção positiva, exclamativo de emoção negativa, etc).

Tabela 2 – Análise linguística da frase

Nível Morfol.	O	marquês	e	Taveira	moviam	lentamente	as	pedras	se	u	pala	co	u	a	de	ran	sur
	Det	Nome	Conj	Nome	Verbo	Adv	Det	Nome	Prep	Det	Nome	Prep	Det	Nome	Prep	Nome	Adj
Sintagmas	Sintagma Nominal			Sint. Nominal	Sint. Verb.	Sint. Adverb.	Sintagma Nominal	Sintagma Preposicional			Sintagma Prepos.	Sintagma Preposicional					
Grupos frásicos	Sujeito				Predicado Simples				Circunstante		Circunstante						
Grupos Prosódicos	Segmento Frásico			Intra-	Segmento Frásico		Intra-	Segmento Intra-Frásico		Segmento declarativo final							

Outro aspecto que importa ressaltar é o fenómeno do foco no enunciado. O foco é um processo de realce obtido intencionalmente por manipulação prosódica, traduzido frequentemente quer num destaque através do aumento ou da descida acentuada das curvas de f0, quer num aumento de intensidade e de duração da palavra ou da expressão. A análise morfológica dá um contributo importante para a determinação do foco do enunciado, através da distinção imediata entre palavras-função e palavras-conteúdo, sendo estas as únicas candidatas a receberem foco prosódico. Além disso, os fenómenos da topicalização e focalização constituem estratégias sintáticas de realce que também são sublinhadas por foco prosódico.

Para a manipulação prosódica do sinal de fala aplicou-se o modelo Fujisaki, já implementado com sucesso em diversas línguas, que consiste em associar a cada grupo frásico um comando de frase com um determinado valor de f0, segundo o grupo/ padrão prosódico em causa. A cada palavra associa-se também pelo menos um comando de acento, que incrementará f0 em cada sílaba tónica. A marcação temporal dos comandos de frase e de acento deriva portanto da análise linguística e da marcação de fronteiras entre sílabas e entre grupos frásicos/prosódicos e, como se pode observar no exemplo simples da Fig. 1, o resultado aproxima-se

bastante do f0 original:

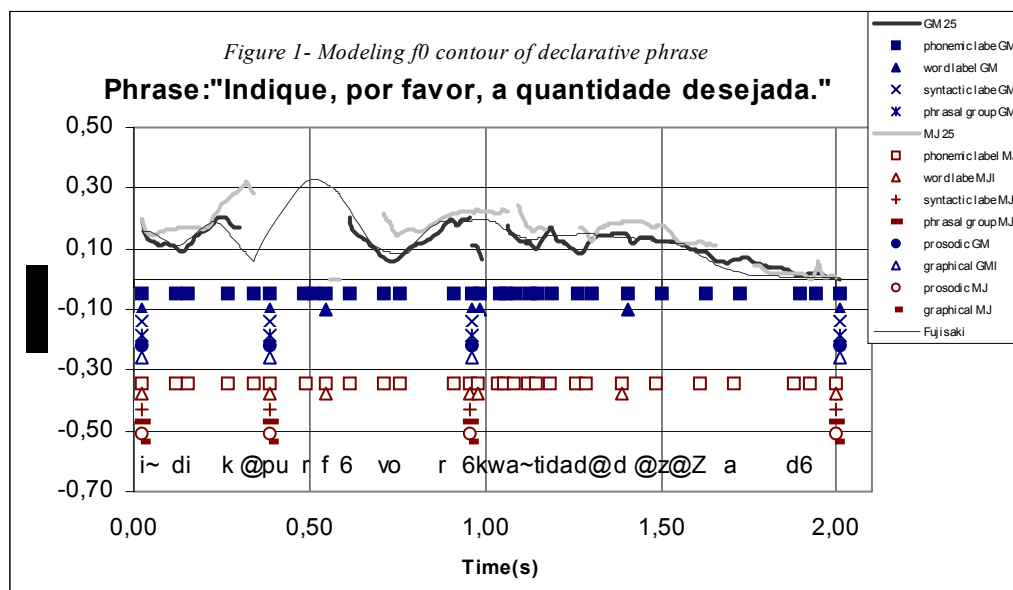


Fig. 1 -

5. CONCLUSÃO

Neste trabalho pretendemos mostrar, na sequência da experiência desenvolvida no LPF-ESI da FEUP/DEEC, de que forma o processamento linguístico computacional automático aplicado à síntese da fala pode recorrer a vários domínios do conhecimento linguístico (fonético, fonológico, morfológico, sintático, semântico e pragmático, entre outros) com o objectivo de construir uma aparelhagem de regras programáveis que dotem o sintetizador da capacidade necessária para a descodificação do texto escrito. A linguística, quando aplicada ao processamento da fala permite construir um sistema, em princípio, menos pesado, uma vez que servindo-se de listas de regras, não necessita de utilizar extensas bases de dados. Por outro lado, um sistema regido por regras mostra-se geralmente mais robusto e atinge em princípio mais facilmente uma menor taxa de erro, do que um sistema baseado em análise estatística.

No presente artigo ilustra-se, essencialmente com a matéria linguística, a algoritmia necessária num sistema de conversão texto-fala desde a entrada de texto até à entrada da unidade de geração de sinal de voz. A implementação dos mecanismos atrás explicados pode ser realizada por meio de programas desenvolvidos em diversas linguagens, destacando-se, pela flexibilidade a linguagem PROLOG e pela eficácia a linguagem C.

6. REFERÊNCIAS

Andradre, A. 1994. 'Sobre a variação fonética de /i/. Uma primeira abordagem', in Actas do X Encontro Nacional da APL, Évora. pp. 25-43.

Delgado-Martins. *Fonética do Português: trinta anos de investigação*. Lisboa, Caminho. 2002. Capítulos 12, 24, 25 e 26.

Freitas, M. J. *Para a aquisição da estrutura silábica do Português Europeu*. Dissertação de Doutoramento, Universidade de Lisboa, 1997.

Freitas, M.J. *A sílaba e os seus constituintes*. Comunicação apresentada na Universidade do

Minho em Fevereiro de 2002, no âmbito do Ciclo de Conferências do Ciclo Linguístico de Braga, 2002.

Frota, S. *Prosody and Focus in European Portuguese*. PhD Thesis. New York/ London, Garland Publishing, 2000.

Jurafsky, D.; Martin, J. *Speech and Language Processing*. New Jersey, Prentice-Hall, Inc. 2000.

Mateus, M.H. *Aspectos da Fonologia Portuguesa*. Publicações do Centro de Estudos Filológicos, Lisboa, 1975.

Monaghan, A. 'Prosody in Synthetic Speech – problems, solutions and challenges', in *Improvements in Speech Synthesis - Cost 258: The Naturalness of Synthetic Speech*. England, John Wiley & sons, LTD, 2001.

Teixeira, J.; Freitas, D. 'Acoustic characterisation of the tonic syllable in Portuguese', in *Improvements in Speech Synthesis - Cost 258: The Naturalness of Synthetic Speech*. England, John Wiley & sons, LTD, 2001.

Vilela, M. *Gramática da Língua Portuguesa*. Coimbra, Almedina, 1999.

VoiceXML - <http://www.w3.org/TR/voicexml20/>

MathML - <http://www.w3.org/TR/MathML2/>

Sable - <http://www.w3.org/TR/voicexml20/>

SSML - <http://www.w3.org/TR/speech-synthesis/>

TPML – consultar os autores.

SAMPA - Speech Assessment Methods Phonetic Alphabet is a machine-readable phonetic alphabet (www.phon.ucl.ac.uk/home/sampa/home.htm).

IPA – International Phonetics Association (<http://www.arts.gla.ac.uk/IPA/ipa.html>).