Gramática sem dicionário: Relatório preliminar

Diana Santos
Carla Fernandes
Rui Marques
José Carlos Medeiros
INESC
R. Alves Redol,9
P-1000 Lisboa
Portugal
{dms,cmf,rprm}@inesc.inesc.pt

Março 1992

Introdução

Este relatório descreve o trabalho realizado para o projecto de síntese de fala a partir de texto, a cargo do Luis Caldas de Oliveira em colaboração com o CLUL[Oliveira et al. 91], cujo objectivo era usar conhecimento linguístico, particularmente no domínio da morfologia e da sintaxe, para melhorar a qualidade da voz sintetizada.

Este documento divide-se em quatro partes, cada uma a cargo de um dos autores, e pela mesma ordem:

1. Arquitectura genérica do sistema;
2. Sintaxe sem dicionário e sua implementação;
3. Fundamentos linguísticos do analisador morfológico;
4. Implementação de um sistema morfológico robusto e genérico.

Finalmente, apresenta-se a avaliação geral do sistema completo à data da escrita deste relatório.

Contents

Chapter 1

Arquitectura

O sistema foi desenhado de forma a que todos os módulos que o constituem fossem utilizáveis, e fizessem sentido, fora da aplicação específica em que se enquadram aqui.

Uma das preocupações foi a de separar estritamente a informação léxica do conhecimento morfológico e sintáctico. Os módulos da morfologia e da sintaxe funcionam pois independentemente de existir ou não um dicionário que classifica as palavras (com excepção das palavras fechadas, cuja lista deve existir).

Assim, se o analisador morfológico não tiver acesso a informação sobre os verbos que existem, deverá contudo indicar que - se tal verbo existir - nos encontramos em presença de tal(is) forma(s).

Da mesma maneira, a sintaxe deverá tratar as palavras que podem ser formas de verbos e tentar averiguar se num dado contexto tais palavras o são ou não, tarefa que é sempre necessária dada a grande frequência com que se verifica a homonímia entre formas de verbos e nomes, por exemplo (veja-se *forma; começo; andas; canto; vão;* etc.).

Estes módulos permitirão depois decidir quais as palavras (ou informação sobre as palavras) que se devem encontrar no léxico, por oposição às que a sintaxe e/ou morfologia permitem identificar inequivocamente. Como tal, este trabalho pode ser considerado como o desenvolvimento de uma ferramenta preliminar para o desenho de um dicionário motivado linguisticamente.

Para o objectivo que nos propusemos tratar — nomeadamente, a detecção dos verbos, finitos ou não — o sistema tem o seguinte desenho global:

- Sistema morfológico (identificado por 1 na figura) com dicionário de palavras fechadas (A na figura), e que constitui um sistema genérico de morfologia do português.
- Pequeno dicionário com algumas palavras, para melhorar o desmpenho do sistema global (como estamos primordialmente interessados nos verbos, resume-se a uma lista de formas verbais que são sempre formas verbais finitas) (C na figura).
- Pequeno tradutor do resultado do analisador morfológico para servir de entrada ao analisador sintáctico, em praticular, classificar com um único nome palavras ambíguas em termos da sua parte da oração (2 na figura).
- Sistema de sintaxe de superfície (3 na figura), cujo objectivo é melhorar a classificação obtida pela morfologia em relação à classificação de partes de oração. Este sistema usa o contexto imediato (a classificação de uma ou duas palavras antes ou depois) para decidir sobre a classificação de uma dada palavra. Usa apenas expressões regulares.
- Sistema de decisão final (4 na figura). Visto que os módulos acima não permitem obter certeza na maioria dos casos, mas apenas maior ou menor probabilidade de uma determinada palavra ser verbo ou não (ou conjunção, ou artigo,...) o sistema final deverá tomar uma decisão heurística. Por não ter havido tempo para desenvolver este componente, o seu funcionamento reduz-se a aceitar como verbos aqueles para os quais tenha havido regras que o proponham e a rejeitar aqueles para os quais tenha havido regras que o desaconselhem.

Figure 1.1: Arquitectura do sistema global

Na figura, B representa um dicionário maior (eventualmente cobrindo mais e mais o léxico, e que poderá ser ou não acedido pelos analisadores morfológico e sintáctico).

Este sistema global de sintaxe poderá e deverá ser melhorado com a incorporação de regras de estrutura sintagmática e de regras que entrem em consideração com a concordância, tarefas que pretendemos implementar no futuro.

Chapter 2

Um analisador sintáctico para reconhecimento de partes da oração

2.1 Introdução

Pretende-se com este trabalho desenvolver um modo de reconhecimento das partes de oração em qualquer frase escrita a analisar para síntese de voz com acesso a um dicionário muito reduzido.

A parte de oração que se pretende reconhecer em primeiro lugar é o verbo, procurando-se identificar prioritariamente as formas finitas, mas estabelecendo-se também uma marcação para os infinitivos pessoais, particípios passados e gerúndios.

Este módulo sintáctico pressupõe o seguinte pequeno dicionário:

O dicionário será composto pelas palavras pertencentes a classes gramaticais fechadas (artigos, pronomes, preposições, conjunções e advérbios) que não sejam homónimas de verbos ou nomes. De notar que só os advérbios não terminados em "-mente" se encontram no dicionário; os advérbios terminados em "-mente" são classificados como verbos possíveis por se poderem confundir com formas de verbos cujo radical termina em "ment" (tais como *mentir* ou *comentar*).

Convém desde já indicar que, ao tentarmos elaborar regras que possibilitem o reconhecimento das partes de oração em qualquer dada frase, partimos do princípio de que as frases não contêm erros gramaticais.

2.2 Descrição do módulo

O módulo que diz respeito à Sintaxe será composto por um primeiro nível de regras de informação parcial que poderão ser garantidas a 100% ou garantidas com uma determinada probabilidade. Num segundo nível, as regras são conjugadas de modo a se escolher o verbo principal caso haja mais do que um numa frase.

As categorias morfológicas possíveis à entrada da análise sintáctica são as seguintes: possível artigo (pa), preposição (pr), pronome pessoal (pn), possessivo (ps), clítico (cl), verbo certo (vc), verbo finito[1] possível (vp), verbo particípio passado possível (vppp), verbo infinitivo (pessoal) possível (vipp), verbo gerúndio possível (vgp), não-verbo (nv), advérbio (adv), advérbio de negação "não" (nao), conjunção (cj), conjunção separadora de orações (cjs), contracção de preposição com artigo (ctr), partícula "se" (se), preposição possível "a" (app), e auxiliares "ter", "ser", "estar" e "ir".

As palavras previamente definidas como verbos no dicionário produzirão sempre a saída V (verbo) nos níveis "Marcador" e "Resultado" (v. secção 2.3).

Outras saídas possíveis naqueles mesmos níveis são as de VP, VPP, VPPP, VIP, VIPP, VG, VGP, SEP (separador de orações) ou F (falso).

[1] Consideramos como verbos finitos aqueles que possuam marca de tempo e pessoa.

As palavras que nunca podem ser verbo (nv), porque não obedecem às terminações das regras 1 ou 2 ou porque a regra assim o determina, terão uma saída F.

2.2.1 Regras sintácticas

As regras a seguir descritas foram criadas com base num corpus composto por 650 frases provenientes dos mais diversos textos [Corpora 90].

"verbo-certeza" : as palavras previamente definidas como verbos no dicionário terão sempre a saída V.

"com-clítico" : um verbo possível (ou infinitivo pessoal possível) seguido de um clítico (unidos por hífen) será sempre identificado como verbo.

"não-verbo" : as palavras que nunca poderão ser verbo, terão uma saída F.

"depois-prep" : depois de preposição não ocorre verbo finito, mas podem ocorrer infinitivos pessoais.

"part-pass" : um particípio passado possível será classificado como VPP, se ocorrer depois de um verbo auxiliar como "ter" ou "ser", conjugados em qualquer tempo e modo, exceptuando-se a forma "tido".

"infinitivo" : um verbo infinitivo pessoal possível será classificado como VIPP , se ocorrer depois de um verbo possível (vp) ou de um outro "vipp", ou ainda depois de uma preposição, pronome ou partícula "se".

"vipp-vp" : antes de um verbo finito não ocorre verbo infinitivo.

"depois-infi" : depois de infinitivo não ocorre verbo finito.

"gerúndio" : depois dos auxiliares "ir" ou "estar" (não conjugados no gerúndio nem no particípio passado), ou depois de "se" ou "não", pode ocorrer um verbo no gerúndio.

"vgp-vp" : depois de gerúndio não pode ocorrer verbo finito.

"depois-pron" : depois de um pronome pessoal pode surgir o verbo.

"depois-poss" : depois de adjectivos possessivos não surge o verbo.

"se-vp" : Quando "se" não se encontra precedido de hífen, o verbo surge depois dele.

"conj-se" : Depois de uma conjunção seguida de "se", pode surgir o verbo.

"se-pn" : Depois de um pronome precedido de "se", pode surgir o verbo.

"se-vipp" : Depois de "se", pode surgir um verbo infinitivo.

"adv-se" : Depois de "se" antecedido de advérbio, pode surgir um verbo finito.

"depois-nao" : Depois de "não", pode surgir o verbo.

"nao-se" : Depois de "se" antecedido de "não", pode surgir o verbo.

"nao-pn" : Depois de pronome precedido de "não", pode ocorrer o verbo.

"se-nao" : Depois de "se" seguido de "não", pode surgir o verbo.

"depois-artigo" : Depois de possível artigo definido ou indefinido não surge o verbo.

"depois-contrc" : Depois de contracção de preposição com artigo não surge o verbo.

2.3 Implementação do módulo

O programa para detectar os verbos consiste num conjunto de regras sintácticas escritas em SCYLA[Lazzaretto 90], uma linguagem de programação criada especificamente para o desenvolvimento de regras texto-para-voz e, em termos gerais, para o processamento linguístico que envolva regras contextuais a vários níveis (e.g. transcrição fonética, controlo da prosódia, fenómenos fonológicos, etc.). As iniciais "SCYLA" provêm de "Speech Compiler (for) Your Language".

O nosso programa possui quatro níveis de representação diferentes, designadamente : "MORFOLOGIA" (o nível de entrada em que as palavras estão classificadas de acordo com a sua categoria morfológica possível); "CONTAGEM" (em que se conta o número de vezes que cada palavra pode ser identificada como verbo); "MARCADOR" (onde as palavras terão já, segundo as regras de escolha criadas, apenas a marcação "V", "VP", "VIP", "VIPP", "VPP", "VPPP", "VG", "VGP", "SEP" ou "F",) e, finalmente, o nível "RESULTADO" (que desambiguará os casos em que exista mais do que um "VP" em cada oração, no nível "marcador", escolhendo apenas um deles.)

2.3.1 Regras funcionais

Apresentam-se, de seguida, as regras que dizem respeito ao modo de funcionamento do programa:

"inicializa-contagem" : inicia-se a possibilidade de cada palavra poder ser contada como verbo sempre que uma regra se lhe aplique.

"alternativa" : sempre que o valor da contagem seja zero, as categorias "vppp", "vp", "vipp" e "vgp" são transferidas para o nível "marcador" (em maiúsculas) para ainda poderem ser processadas no nível seguinte.

"máximo" : o valor maior do nível "contagem" é guardado na variável "máximo" que ganha um novo índice sempre que surge um separador (SEP), uma vez que se inicia outra oração.

"intermédio" : copia do nível "marcador" para o "resultado" para neste nível se fazer a selecção dos vários VPs que ocorram.

"final" : entre vários VPs com um determinado índice proveniente da contagem, será escolhido como verbo mais provável aquele que tenha o índice mais elevado (máximo). Sempre que surja um SEP (separador de orações), a variável máximo ganha um novo índice e a contagem é reiniciada independentemente do valor máximo anterior. Quando surge um ponto final, o índice da variável "máximo" volta a ser zero, assim como os valores dos próprios máximos, de modo a que o programa esteja pronto a processar a nova frase.

2.4 Padrões relativos às regras sintácticas

Apresentam-se abaixo os padrões sintácticos das regras já implementadas. Os números referem-se aos resultados de testes efectuados sobre as primeiras 221 frases do corpus.

As categorias em itálico são aquelas sobre as quais recai o teste em cada regra.

A sigla "-VP" significa "VP negativo", ou seja, provavelmente falso.

- *vp* cl (Certo)
- *vipp* cl (Certo)
- pr *vp* (-VP) (118 instâncias - 0 problemas)
- ctr *vp* (-VP) (113 instâncias - 0 problemas)
- pa *vp* (-VP) (110 instâncias - 3 problemas)
- pn *vp* (91 instâncias - 41 problemas)

- pai *vp* (-VP) (67 instâncias - 0 problemas)
- app *vp* (-VP) (60 instâncias - 0 problemas)
- pr *vipp* (41 instâncias - 5 problemas)
- vp *vipp* (38 instâncias - 9 problemas)
- ps *vp* (-VP) (36 instâncias - 0 problemas)
- vipp *vp* (-VP) (36 instâncias - 2 problemas)
- *vipp* vp (-VP) (36 instâncias - 0 problemas)
- se *vp* (27 instâncias - 0 problemas)
- app *vipp* (26 instâncias - 1 problema)
- nao *vp* (20 instâncias - 2 problemas)
- pn *vipp* (8 instâncias - 1 problema)
- se *vipp* (5 instâncias - 1 problema)
- vipp *vipp* (5 instâncias - 2 problemas)
- se *vipp* (5 instâncias - 0 problemas)
- adv se *vp* (5 instâncias - 0 problemas)
- vgp *vp* (-VP) (4 instâncias - 0 problemas)
- cj se *vp* (4 instâncias - 0 problemas)
- ir *vgp* (3 instâncias - 0 problemas)
- ter *vppp* (2 instâncias - 0 problemas)
- se pn *vp* (2 instâncias - 1 problema)
- nao se *vp* (2 instâncias - 0 problemas)
- ser *vppp* (1 instância - 0 problemas)
- estar *vgp* (1 instância - 0 problemas)
- nao pa *vp* (1 instância - 0 problemas)
- nao pn *vp* (1 instância - 0 problemas)
- se *vgp* (0 instâncias)
- se nao *vp* (0 instâncias)
- nao *vgp* (0 instâncias)
- pa *vgp* (-VP) (0 instâncias)
- se, *vp* (-VP) (0 instâncias)
- se adv *vp* (0 instâncias)

2.5 Exemplos de regras implementadas no programa

2.5.1 REGRA preposição

Imediatamente depois de preposição não pode surgir um verbo finito. Exemplos:

- "...teria o efeito de *tornar*..."
- "Existem grandes dificuldades em *obter*..."
- "...depois de *amanharem* o linho..."

REGRA:

$$vp- > F$$
$$/pr...^{2}$$
$$/app...;$$

Um "vp" precedido de preposição terá a saída F.

Para testar esta regra escolheram-se as preposições "a", "de", "em" e "para".

Da análise de todos os contextos em que a preposição "de" ocorre no corpus, verificou-se que todos os casos são correctamente analisados ao aplicarem-se as regras acima descritas.

2.5.2 REGRA enclíticos

Os enclíticos encontram-se sempre junto a verbos. Exemplos:

- "O estudo dos sons...chama-*se*, em gramática,..."
- " ...alargar-*se*-ia cada vez mais..."
- "...vai-*te* lixar, esta' bem?"

REGRA:

$$vp- > V$$
$$/...cl;$$

Um "vp" seguido de um clítico terá a saída V.

2.5.3 REGRA possessivo

Depois de um adjectivo possessivo não surge o verbo.

REGRA:

$$vp- > F$$
$$/ps...;$$

Esta regra é confirmada a 100% no corpus, uma vez que ali não existem exemplos em que os possessivos sejam pronomes. Nesse caso poder-se-ia seguir o verbo ao possessivo, mas, de qualquer forma, a frequência de uso dos possessivos como pronomes parece ser muito menor do que como adjectivos e, como tal, pensamos que a regra dará resultados correctos na maioria dos casos.

[2] Deverá entender-se o sinal "/" como o introdutor do contexto da regra descrita, i.e., a condição necessária para que ela seja aplicada. As reticências equivalem ao membro esquerdo da regra descrita antes do contexto.

2.5.4 REGRA part.passado

Os particípios passados possíveis serão classificados como VPP, se ocorrerem depois dos verbos auxiliares "ter" ou "ser" conjugados em qualquer tempo ou modo, exceptuando-se a forma "tido".

REGRA:

$$vppp- > VPP$$
$$/tervp...$$
$$/tervc...$$
$$/tervg...$$
$$/ < ser > ...;$$

2.5.5 REGRA gerúndio

Os verbos gerundivos serão classificados como VGP, se ocorrerem depois dos auxiliares "ir" ou "estar" ou depois das palavras "se" e "não".

REGRA:

$$vgp- > VGP$$
$$/irvc...$$
$$/irvp...$$
$$/estarvc...$$
$$/estarvp...$$
$$/se...$$
$$/nao...;$$

2.5.6 REGRA infinitivo

Os verbos infinitivos pessoais possíveis ("vipp") serão classificados como VIPP se ocorrerem depois de um "vp" ou de um outro "vipp", ou ainda depois de uma preposição, pronome ou partícula "se".

REGRA:

$$vipp- > VIPP$$
$$/vp...$$
$$/vipp...$$
$$/pr...$$
$$/app...$$
$$/pn...$$
$$/se...;$$

Os seguintes casos do corpus são incorrectamente analisados, uma vez que as palavras assinaladas a itálico se confundem com verbos infinitivos:

- 34: "...quarto de *dormir*..."
- 53: "...vem de *par* com..."
- 69: "...a maior quantidade possível de *ar*..."
- 85: "...com *ar*..."
- 98: "...espectáculo de beleza e de *horror*..."
- 151: "...de *ares*..."

- 272: "...uma espécie de *limiar*..."
- 273: "...uma permanente fonte de *prazer*..."
- 345: "...localidades de *maior* interesse..."
- 498: "...pertencer à equipe de *mar*."
- 638: "...percentagem de *valor* acrescentado..."
- 297: "...com a Casa Pia em *particular*."
- 100: "...a *partir* da abertura do nosso século."
- 641: "...estar a *par* da tecnologia..."

2.5.7 REGRA separador

As conjunções "embora, mas, pois, porque e quando" marcarão o fim de uma oração e o início de outra.

Note-se que algumas conjunções são coincidentes com advérbios (caso de "embora") ou com formas verbais (caso de "como").

No entanto, no nosso corpus, as palavras "embora" e "como" ocorrem sempre como conjunção.

REGRA:

$$cjs- > SEP;$$

2.5.8 REGRA "não"

Depois do advérbio de negação "não", pode surgir o verbo:

REGRA:

$$vp- > VP$$

$$/nao...;$$

A palavra "não" ocorre em 111 frases do corpus, das quais se escolheram 11 pelas suas diversas estruturas sintácticas.

Contextos em que "não" ocorre no corpus:

- nao vc pa
- nao vc vip
- nao se vp
- se nao vp vip
- pn nao pa/pn vp
- pr nao pn vip
- nao pa
- nao nv
- nao adv
- nao vip
- nao pr

2.5.9 REGRA "se"

Quando a partícula "se" não ocorre depois de hífen, o verbo pode surgir depois dela.

REGRA:

$$vp- > VP$$

$$/se...;$$

Das frases testadas, todas são correctamente analisadas, pois, nos casos em que não é o verbo que se segue a "se", são palavras pertencentes às classes gramaticais previamente definidas.

Ocorrendo em 104 frases do corpus, verificaram-se 12 contextos diferentes de ocorrências de "se":

- Se pr vip
- adv se pn
- se pn vp
- se pa/pn vp
- se, vp nv
- se adv(bem) que
- cj se vp
- se adv vp
- se nao vp
- se vipp
- se vp vp
- se pa pn

2.6 Estudos para a distinção artigo/pronome

(Esta secção foi escrita em colaboração com o RUI MARQUES.)

Uma vez que as palavras "o, os, a, as" podem ser artigo definido ou pronome, apresentam-se as seguintes conclusões que, apesar de não terem ainda sido suficientes para a criação de regras sintácticas, poderão eventualmente, de futuro, contribuir para desambiguar alguns dos casos em que aquelas palavras se confundem:

- Na sequência "(Pronome ou artigo) + verbo? + (adjunto, nova frase ou infinitivo)", a primeira palavra será artigo se não existir nenhum dos 3 últimos elementos, mas se algum deles existir não se pode garantir que aquela palavra seja pronome.
- "o, a, os, as" são pronome se forem precedidos das seguintes palavras: alguém, ninguém ou quase. (Não existem exemplos no nosso corpus.)
- As palavras "ora" e "ou" só podem preceder pronome nos contextos "ora...ora" e "ou...ou".
- As palavras "jamais", "nunca" ou "não" precedem pronome se a este se seguir outra palavra que possa ser verbo seguida ou não de advérbio ou infinitivo.
- As palavras "se", "quando" e "onde" precedem pronome se a este se seguir outra palavra (o verbo?) e um sinal de pontuação diferente de vírgula.
- A palavra "quase" precede pronome se a este se seguir outra palavra (o verbo?) e eventualmente uma conjunção ou um infinitivo.
- As palavra "porque" e "que" precedem pronome se a este se seguir outra palavra seguida de um infinitivo.

- A palavra "sempre" só pode preceder artigo quando segue o verbo. Assim, se na frase já houver uma palavra candidata a verbo e "sempre" ocorrer no contexto "vp + sempre + o, a, os, as", estes serão artigo independentemente do contexto que se segue. (Ex.:331: "... a sua opinião nos merece sempre o maior respeito.")

- A palavra "ainda" precede pronome se a este se seguir outra palavra (o verbo?) e um infinitivo. (Não há exemplos no corpus.)

- Se "o, a, os, as" precederem maiúscula, então são artigo.

Contextos em que "o" ocorre no corpus:

- pa/pn vp
- pa/pn nome
- pa/pn que
- pa/pn ps
- pa/pn vip

2.7 Problemas

- Os advérbios terminados em "-mente" são classificados como verbos possíveis por se poderem confundir com formas de verbos cujo radical termine em "-ment". Esta questão poderá ser resolvida com uma listagem mais exaustiva dos advérbios no dicionário.

- O facto de o programa detectar demasiados "VPs" no último nível de saída ("Resultado") deve-se à impossibilidade de, neste momento, se definir o número de orações por frase. Uma hipótese de resolução deste problema parecia ser a da utilização das conjunções como separadores de orações, mas apenas o conseguimos com cinco delas (classificadas no programa como "cjs").

- A frase "Quantos terminais se lhe poderão ligar?" é um exemplo de frase impossível de analisar apenas através da sintaxe, pois o sintagma "Quantos terminais" poderá ter função sintáctica de sujeito ou de objecto directo e a partícula "se" a de sujeito ou de reflexo. Para uma análise correcta seria necessário recorrer também à semântica.

Chapter 3

Bases linguísticas do analisador morfológico

3.1 Introdução

O objectivo de um analisador morfológico é atribuir a cada qualquer palavra da língua a(s) categoria(s) morfológica(s) a que possa pertencer, bem como a informação gramatical associada a cada forma (por exemplo informação de *número, pessoa e tempo* para as formas verbais e de *número e género* para as formas nominais ...). No presente caso pretende-se que a análise morfológica seja feita evitando o recurso ao dicionário. Este contém apenas as palavras que formam as classes fechadas (*preposições, advérbios, pronomes e conjunções*) e um pequeno número de formas verbais de uso bastante frequente. O facto de não se usar dicionário justifica a criação de um conjunto de regras heurísticas. Neste sistema, essas regras consistiram no levantamento de terminações de radicais impossíveis para certas conjugações (ver *anexo 1*). Foi, por exemplo observado que não há radicais de verbos da segunda conjugação que terminem em *f*. Estas regras permitem reduzir a informação gramatical supérflua associada a cada forma verbal.

3.2 Regras da Morfologia

Num sistema que parta da análise de um texto escrito, o analisador morfológico pode fazer uso de dois tipos de informação: regras da grafia de uma língua (exemplo: *palavras ligadas por* hífen*, que não sejam clíticos formam uma palavra composta*), e terminações de cada classe morfológica. Há terminações que são exclusivas de uma classe morfológica e outras que são comuns a várias classes. Há por isso que distinguir entre *terminações certas* e *terminações possíveis*. As terminações certas de *verbo* (ver anexo 2) são exclusivas da classe dos *verbos*, assim como as terminações certas de *nome/adjectivo* (ver anexo 3) não podem ser terminações de *verbo*. Formam, portanto, conjuntos disjuntos. Por seu lado, o conjunto das terminações possíveis de *verbo* (ver anexo 2) são também terminações possíveis de *nome/adjectivo*. Assim, ao analisar uma palavra que termine por exemplo em *as*, a morfologia deve indicar a possibilidade de essa palavra ser *verbo* ou *nome/adjectivo*, mesmo que não se trate de palavras homónimas. Isto porque tanto existem formas verbais que terminam em *as* (exemplo: *corras, controlas ...*) como formas nominais com a mesma terminação (exemplo: *camisas, crianças ...*).

3.3 A classe dos Nomes/Adjectivos

Não se fez a separação entre a classe de *nomes* e a classe *adjectivos* porque a maioria das palavras que podem ser *adjectivo* podem também ser *nome*. Seria então necessário criar três listas em vez de uma única (: a lista das terminações que só podem ser nome, a lista das terminações que só podem ser adjectivo, e a lista das terminações que podem ser nome ou adjectivo mas não verbo).

Optou-se por se criar uma única lista que contivesse as terminações de *nomes* ou *adjectivos* que não são terminações possíveis de verbo. Cabe à análise sintáctica fazer a desambiguação.

3.4 Metodologia

Todas as terminações de verbo são portadoras de informação gramatical (indicam tempo, modo, pessoa e número). Logo, qualquer novo verbo que entre na língua, obedecerá obrigatoriamente a essas terminações. O mesmo não se pode dizer das terminações que podem ser de *nome* ou de *adjectivo* mas não de *verbo*, as quais constituem um conjunto bastante mais numeroso. As terminações de *nome/adjectivo* listadas em anexo não são um levantamento exaustivo de todos os *nomes/adjectivos* actualmente existentes na língua. Mesmo que o levantamento tivesse sido baseado num dicionário que contivesse todas as palavras portuguesas dessa classe, haveria a possibilidade de se criar (ou importar) um neologismo cuja terminação fosse diferente de todas as outras já existentes. A quantidade de terminações desta classe a considerar é, portanto, o resultado de uma escolha. Não há, como para os verbos, um pequeno número de terminações exclusivas desta classe.

Igualmente arbitrário é o número de letras que compõem as terminações. Neste sistema esse número foi fixado num máximo de quatro, ou seja, consideram-se terminações com uma, duas, três e quatro letras. Quanto maior for o número de letras que se considere, maior será o número de palavras abrangidas e consequentemente a percentagem de sucesso da análise morfológica. Por outro lado, mais espaço ocupará o programa. Essa escolha não obedece, portanto a um critério científico, é o produto de uma decisão de implementação.

3.5 Implementação

À data da escrita deste relatório apenas foi implementada a informação linguística relativa a formas verbais. Falta implementar a informação relativa a *nomes/adjectivos* e o reconhecimento de advérbios em *mente* (que por formarem uma subclasse não fechada não vêm listados no dicionário actual).

3.6 ANEXO 1 - Restrições à explosão morfológica

Uma forma de refinar a análise de uma forma verbal consiste em criar regras heurísticas que restringem o número de lemas possíveis.

3.6.1 Verbos da segunda conjugação

Quanto aos verbos da segunda conjugação (terminados em *e*), podemos observar que os seus lemas não podem ter as seguintes terminações (à frente de cada uma é indicado, como exemplo, um verbo que não existe, mas que é actualmente dado como raíz possível de uma forma verbal):

- e (*romanceer)
- a (*distraer)
- f (*alcatifer)
- i (*associer)
- n (*iluminer)
- ib (*iliber)
- rc (*esforcer)
- uc (*baloucer)
- nh (*ganher)
- uv (*viuver)
- uz (*cruzer)

Podemos restringir ainda mais os lemas possiíveis de verbos terminados em *er* se acrescentarmos as seguintes terminações:

- com a terminação *der* apenas existem verbos terminados em:
  1. eder
  2. nder
  3. oder
  4. rder
- com a terminação[1] *ler*
  1. aler
  2. eler
  3. oler
- com a terminação *mer*
  1. emer
  2. omer
- com a terminação *oer*
  1. doer
  2. moer
  3. roer
  4. soer
- com a terminação *per*, a única possibilidade é a forma:
  1. -romper

[1] A própria terminação pode constituir um verbo (*ler, ser, ter, ver*). As regras foram feitas apenas para os casos em que a terminação se junta a outras letras.

- com a terminação *rer*:
    1. crer
    2. erer
    3. frer
    4. rrer
- com a terminação *ser*:
    1. oser
    2. sser
- com a terminação *ter*
    1. ater
    2. eter
    3. nter
    4. rter
    5. ster
- com a terminação *uer*
    1. guer
    2. quer
- finalmente com a terminação *xer* os únicos verbos são os que terminam em:
    1. mexer

3.7 Verbos da terceira conjugação

A mesma metodologia se pode seguir para os verbos da terceira conjugação (terminados em *ir* no infinitivo impessoal). Assim, podemos listar como terminações impossíveis de lemas de verbo da terceira conjugação as seguintes:

- i (*piir)
- e (*romanceir)
- f (*alcatifir)
- o (*doir)
- x (*mexir)
- od (*acomodir)
- eg (*elegir)
- ch (*enchir)
- lh (*encolhir)
- gn (*designir)
- on (*condicionir)
- sn (*grasnir)
- cr (*massacrir)
- fr (*decifrir)
- lr (*galrir)
- nr (*honrir)
- vr (*livrir)
- lt (*faltir)

- ot(* boicotir)
- pt (*optir)
- xt (*pretextir)
- du (*graduir)
- iv (*cultivir)
- lv (*salvir)

Com as terminações que se seguem, apenas são possíveis as formas indicadas:

- b
    1. ibir
    2. mbir
    3. subir
- c
    1. ressarcir
    2. aducir
- l
    1. balir
    2. falir
    3. elir
    4. olir
    5. ulir
- m
    1. bramir
    2. dormir
    3. emir
    4. imir
    5. umir
- p
    1. lpir
    2. carpir
    3. despir
    4. cuspir
    5. upir
- s
    1. entresir
    2. transir
    3. tossir
- ld
    1. gualdir
- nh
    1. renhir
    2. grunhir

- an
    1. banir
- ar
    1. parir
- dr
    1. apodrir
- gr
    1. denegrir
- at
    1. latir
- cu
    1. imiscuir
- su
    1. possuir
- zu
    1. zuir
- ev
    1. sobrevir
- uv
    1. ouvir
- nv
    1. convir

  item ov
    1. provir

3.7.1 Verbos da primeira conjugação

A primeira conjugação (verbos terminados em *ar* no infinitivo impessoal), sendo a mais produtiva, não manifesta praticamente qualquer restrição às terminações do lema. A única restrição encontrada é a terminação *moar*. Não há nenhum verbo cujo lema termine em *mo*, que não pertença à segunda conjugação.

3.8 ANEXO 2 - Identificação de formas verbais pela morfologia

Este anexo foi escrito em colaboração com a CARLA FERNANDES.

3.8.1 Verbos certos

Serão identificadas como verbos as palavras que surjam junto a enclíticos (-a -as -o -os -me, -ma(s), -mo(s), -te, -ta(s), -to(s), -se, -lhe(s), -lha(s), -lho(s), -na(s), -no(s), -lo(s)) ou as que tenham as seguintes terminações:

- "s": -ámos
- "m": -am -ssem
- "u": -ou -iu

3.8.2 Verbos possiíveis

Outras terminações sem as quais as palavras não podem ser identificadas como verbos (apesar de muitas delas se confundirem com as de outras palavras):

Verbos finitos

- "s": -ais -as -eis -es -ias -mos -rás
- "r": -ar -er -ir -or
- "m": -em
- "z": -az -ez -iz
- "i": -ei -rei
- -a -á -e -i -o

Particípios passados

- -ado
- -ido
- -ído
- -ito
- -sto
- -rto

Gerúndios

- -ando
- -endo
- -indo
- -ondo

Infinitivos pessoais

- -ar -er -ir
- -ares -eres -ires
- -rmos
- -rem

Verbos impossíveis

Palavras iniciadas por maiúsculas não podem ser verbo, excepto se se encontrarem em início de frase.

Não podem além disso ser verbo todos as palavras que ou não possuam as terminações listadas acima ou que possuindo-as, estejam cobertas pelas restrições à explosão morfológica listadas no Anexo 1.

Formas previamente definidas como verbos

- é
- está
- foi
- há
- tem
- pode
- existem
- poderá
- começou
- faz
- podem
- têm
- temos
- tenho
- vai
- começava
- considerei
- deverá
- fazem
- tornam

3.9 ANEXO 3 - Identificação de outras classes morfológicas

3.9.1 Palavras que só podem ser nome ou adjectivo

- qualquer letra isolada pode ser nome ou adjectivo (exemplo: *alínea a*)
- palavra com dois hífens que não liguem clíticos é nome
- palavra iniciada por maiúscula, que não esteja em início de frase é nome. Se estiver em início de frase pode não ser.
- sequência de palavras iniciadas por letra maiúscula é nome.
- palavra iniciada por letra maiúscula + preposição + palavra iniciada por letra maiúscula é nome

3.9.2 Terminações de palavras que só podem ser nome ou adjectivo

- qualquer palavra cuja última letra tenha til ou acento circunflexo
- palavra terminada em til ou acento circunflexo seguida de *s* (excepção: *pôs, vês, crês, dês*)
- -l
- -n/s (excepções: *vens, tens*)
- -t/s
- -x/s
- qualquer palavra acabada em *z/es* excepto as que acabam em *fez*, *fiz* ou em *fazes*, que são terminações verbais (*fez* pode também ser nome e *refez* pode ser verbo ou adjectivo).
- -ea/s
- ãe/s
- õe/s (excepção: *-põe*s - *pode ser terminação de verbo*)
- ão/s (excepção: *-rão - pode ser terminação de verbo*)
- -xi/s
- -hm/s
- -om
- -um
- -eo/s
- -or/es (excepções: palavras terminadas em *por/pores*, que podem ser infinitivos. São nomes ou adjectivos as palavras: *vapor, tepor, sopor, torpor, estupor*. A palavra *dispor* pode ser nome ou verbo)
- -ur/es
- -ís/es
- -os (excepção: *-mos* - pode ser terminação de verbo)
- -us (excepção: *adeus*)
- -ps
- -au/s
- -éu/s
- -qualquer consoante seguida de *u*/s
- -oba/s
- -sba/s
- -uga/s
- -gla/s

- -mna/s
- -eoa/s
- -moa/s
- -fta/s
- -aua/s
- lua/s
- nua/s
- -oxa/s
- -zua/s
- -rza/s
- -rbe/s
- -cne/s
- -fne/s
- -oxe
- -rze/s
- -uei/s
- boi/s
- -ipi/s
- -cri/s
- -iri/s
- -oti/s
- -lbo/s
- -glo/s
- -tmo/s
- -quo/s
- -oxo/s
- -uxo/s
- -fer/es
- -fir/es
- -xir/es
- -ois
- xis
- -çaba/s
- -iaba/s
- jaba/s
- -oaba/s
- taba/s
- -xaba/s
- -leba/s
- -meba/s
- -reba/s
- giba/s
- -miba/s

- siba/s
- viba/s
- -orba/s
- buba/s
- juba/s
- -caca/s
- faca/s
- iaca/s
- -iaça/s
- jaça/s
- -maca/s
- -oaca/s
- -taça/s
- -uaca/s
- -uaça/s
- -vaça/s
- -heca/s
- -jeca/s
- -leça/s
- -ueca/s
- -veca/s
- -aica/s
- -miça/s
- -undo/s
- -idade/s

O verbo *dar* é o único da primeira conjugação que tem apenas uma letra no radical. Todas as outras palavras com três letras que terminem em *ar* são nomes ou adjectivos.

Chapter 4

Arquitectura e implementação do analisador morfológico

Chapter 5

Avaliação

5.1 Descrição do corpus de teste

O nosso corpus[Corpora 90] é composto por 650 frases retiradas ao acaso de livros, revistas, gramáticas, jornais, textos publicitários, e outros textos diversos[Corpora 90].

Para testar o programa foram usadas as primeiras 221 frases, num total de 6775 palavras.

O dicionário criado para o programa contém 256 palavras, das quais 20 são verbos certos (vc) e 236 pertencem a classes gramaticais fechadas.

Apresenta-se abaixo o número de ocorrências dos verbos e das palavras gramaticais nas 221 frases referidas:

Classificação morfológica	Num. ocorrências
verbos possíveis	2027
verbos certos	106
verbos gerund. possíveis	32
verbos part.pass. possíveis	81
verbos infinitivos possíveis	230
não-verbos	1347
advérbios	227
possíveis artigos def. (sem "a")	314
possíveis artigos ind.	162
conjunções	380
contracções	394
pronomes pessoais	377
adjectivos possessivos	82
preposições (sem "a")	645
palavra "a"	224
palavra "se"	91
palavra "não"	73

5.2 Descrição dos resultados obtidos

Seguem-se os resultados obtidos com as regras morfológicas e sintácticas:

- Verbos possíveis e certos: 2133 - Detectam-se 251 e anulam-se 544.
- Verbos possíveis: 2027 (376 reais[1]) - Detectam-se 153 "VPs" e 544 "-VPs"(provavelmente falsos).
- Verbos certos: 106

[1]Entenda-se por "reais" as formas verbais que tenham realmente a função sintáctica de verbo no corpus.

- Verbos gerundivos possíveis: 32 (29 reais) - Detectam-se 4.
- Verbos possíveis no particípio passado: 81 (74 reais) - Detectam-se 3.
- Verbos infinitivos (pessoais) possíveis: 230 (165 reais) - Detectam-se 108.
- Não-verbos (impossíveis segundo as regras morfológicas) : 1347

5.3 Três possíveis avaliações

Partindo de 482 verbos reais, o módulo da morfologia permite detectar 106 verbos certos e classificar 2027 verbos como possíveis. Destes 2027 verbos possíveis, o módulo da sintaxe detecta 153 como mais prováveis (VP) e 544 como menos prováveis (-VP), ou seja, provavelmente falsos. 1330 verbos não são analisados, o que pensamos vir a ser eventualmente melhorado com a implementação, numa fase posterior, de um terceiro módulo que permita escolher por heurística apenas um VP sempre que em cada oração se obtenha um resultado de mais do que um VP.

Partindo de 2133 verbos, dos quais apenas 482 são verbos reais (VP + VC), o programa detecta 251 (VP + VC) e não dá informação acerca de 1330 verbos.

Dos 482 verbos reais, o dicionário e as regras morfológicas permitem detectar 106 verbos certos. Uma vez que 544 verbos são anulados (como provavelmente falsos), o resultado de verbos possíveis (VP) é de 1483.

Bibliography

[Corpora 90] Corpora. G.C.IBM-INESC, relatório interno INESC. Dezembro 1990.

[Lazzaretto 90] Lazzaretto, Stefano. SCYLA language reference manual. CSELT, Torino, Italy. May 1990.

[Oliveira et al. 91] L.C. Oliveira, M. Céu Viana, I.M. Trancoso. Dixi - Portuguese Text-to-Speech System. *Proceedings of Eurospeech 91 - 2nd. European Conference on Speech Communication and Technology*, Genova, 1991.