Normalização de itens lexicais baseada em sufixos

Marco Gonzalez, Daniela Toscani, Letícia Rosa, Rita Dorneles, Vera L. S. de Lima

PUCRS - Faculdade de Informática
Av.Ipiranga, 6681 – Prédio 16 - PPGCC
90619-900 Porto Alegre, Brasil

{gonzalez, vera}@inf.pucrs.br

Abstract. *The NILP system executes a morphological normalization of lexical items in Portuguese language texts. The system receives a morphological tagged text and adds to each word its canonical form. The verbs (except participles) are reduced to infinitive, and other variable words are reduced to singular masculine form. NILP reads suffix bases and implements deterministic finite automata which are specialized in types such as adjectives, articles, numeral, pronouns, nouns and verbs. NILP prototype reaches up to a 90% precision. A relevant contribution of this work is the construction of the suffix bases. These are adaptable to other purposes (for instance, nominalization process), or they may help morphological normalization of other languages.*

Resumo. *O sistema NILP é um normalizador morfológico de itens lexicais para o Português. O sistema recebe um texto etiquetado morfologicamente e acrescenta, a cada palavra, a sua forma canônica, ou seja, reduz os verbos (exceto particípios) à forma infinitiva e as outras palavras variáveis ao singular masculino. São utilizadas bases de sufixos que são lidas para implementar autômatos finitos determinísticos especializados em adjetivos, artigos, numerais, pronomes, substantivos e verbos. O sistema, em fase de protótipo, alcança uma precisão acima de 90%. Uma importante contribuição deste trabalho é a construção das bases de sufixos. Estas podem ser adaptadas para outras finalidades, como o processo de nominalização, ou podem ser construídas para normalização morfológica de outros idiomas.*

Palavras-chave*: processamento da linguagem natural, morfologia, autômatos finitos, normalização lexical.*

1 Introdução

A normalização morfológica de itens lexicais é o processo que reduz as variações de uma palavra a uma forma única. Este processo é importante para diversas aplicações, entre as quais estão a extração de informação, a sumarização e a classificação de textos, a recuperação de informação e diversas outras. Nestas aplicações, um conceito tratado no texto é tão ou mais importante que o próprio item lexical que veicula tal conceito. A normalização morfológica segue o princípio que estabelece que itens que significam o mesmo conceito tenham a mesma representação de significado [Jurafsky2000].

A normalização morfológica mais usual é o *stemming* (por exemplo, ver [Orengo2001]). Nesta concepção, tradicionalmente há uma redução das palavras ao *stem* (que pode ser entendido como base ou radical).

Em nossa abordagem, optamos por realizar a normalização obtendo a forma canônica[1] [Arampatzis2000], que consiste em uma redução menos drástica da palavra original, quando comparada ao *stem*. Por exemplo, as palavras construções e construiremos são transformadas, por um *stemmer*, em uma mesma cadeia: constru. Por outro lado, a forma canônica mantém a categoria morfológica original e teríamos, para as palavras exemplificadas, respectivamente, construção e construir.

Com tal objetivo, nossa estratégia utiliza a informação que um sufixo agrega à palavra à qual é concatenado. Esta informação é o principal indício que utilizamos para chegar à forma canônica da variante analisada. Nosso sistema de normalização (NILP – Normalizador de Itens Lexicais em Português) recebe, como entrada, um texto etiquetado com as categorias morfológicas de cada palavra, e justapõe a cada uma das palavras da entrada, sua forma canônica. Para partir do estado inicial (a variante da palavra analisada) ao estado final (a forma canônica correspondente), são utilizados autômatos finitos determinísticos especializados nas diferentes categorias morfológicas.

Este artigo se organiza em 4 seções, após esta introdução. A seção 2 apresenta a estratégia do sistema: a abordagem baseada em autômatos finitos determinísticos; a seção 3 descreve as bases de sufixos; a seção 4 apresenta a avaliação inicial de um protótipo do sistema; e a seção 5 tece algumas considerações sobre o trabalho realizado.

2 Estratégia

2.1 Método

Encontramos como componentes formadores de palavras, entre outros, o radical e os afixos [Cegalla1998]. O radical é o elemento básico e significativo, enquanto que os afixos podem ser incorporados a uma palavra (ou ao radical de uma palavra) para formar outra, sendo acrescentados antes (como prefixos) ou após (como sufixos) a mesma. Ao serem inseridos, os sufixos afetam significativamente a palavra original, podendo, inclusive, alterar sua categoria morfológica [Lima1998, Sacconi1999]. Com estas características, os sufixos são utilizados, neste trabalho, para guiar a normalização através da estratégia que é descrita a seguir e apresentada na Figura 1.

Nosso sistema recebe um texto etiquetado com categorias morfológicas e cada palavra, com sua respetiva etiqueta, é tratada através de um autômato especializado na categoria morfológica identificada. É pesquisado o sufixo em uma das seguintes bases de sufixos:

1) Adjetivos (incluindo adjetivos e verbos no particípio),
2) Artigos (definidos e indefinidos),
3) Numerais (cardinais e ordinais),
4) Pronomes (pessoais, demonstrativos, possessivos, indefinidos e relativos),
5) Substantivos e

[1] A forma canônica corresponde ao infinitivo, para os verbos, e ao singular masculino, para as outras palavras variáveis (adjetivos, artigos, numerais, pronomes, substantivos e verbos no particípio).

6) Verbos (exceto os verbos no particípio).

Os verbos no particípio são tratados na base para adjetivos, por terem um comportamento semelhante a estes, quanto à normalização morfológica.

Também é oferecido um tratamento de exceção, em caso de palavras que configurem tal situação.

Tendo estas bases como apoio, a aplicação dos autômatos finitos determinísticos ocorre nos módulos “Reconhece Sufixo”, “Exclui Caracteres” e “Inclui Caracteres” (ver Figura 1).

Após o reconhecimento do sufixo (ou após o tratamento da exceção), é definida a ação a ser tomada:

1) inclusão de caracteres,
2) exclusão de caracteres,
3) exclusão seguida de inclusão de caracteres, ou
4) nenhuma delas.

Há casos em que, para simplificar o número de alternativas exigidas na base de sufixos, após ser executada alguma alteração, repete-se o processo, com o reconhecimento do sufixo alterado. Este é o caso, por exemplo, das palavras com sufixo “inhas”: primeiramente é retirado o “s” e, posteriormente, o sufixo “inha” é tratado. Assim, para algumas palavras, a normalização é otimizada em duas etapas que podem ser, por exemplo, do plural para o singular e, após, do feminino para o masculino.

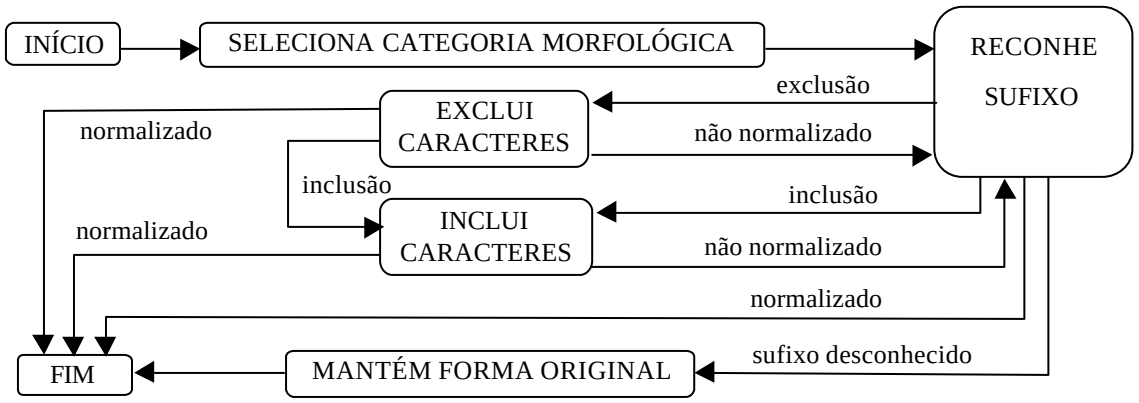

Figura 1. Estratégia de normalização morfológica

Quando um sufixo não é reconhecido, o item lexical é mantido em sua forma original, em princípio porque o sufixo não está presente na base, mas também para atender ao caso dos nomes próprios.

As palavras invariáveis, como as preposições e as conjunções, são mantidas na forma original sem passar pela análise dos autômatos.

2.2 Justificativa das escolhas realizadas

Foi adotada a abordagem com autômatos por ser adequada ao objetivo da normalização morfológica, conforme é explicado a seguir. Um autômato finito incorpora um número predeterminado de estados e um conjunto de transições [Cohen1997]. Em um autômato finito determinístico, haverá um único estado posterior para cada par formado pelo estado atual e por uma transição. O estado atual é representado, neste processo de

normalização, pelo conjunto de caracteres reconhecidos ou alterados até o momento. Uma transição consiste no reconhecimento do próximo caractere ou na execução de uma ação de exclusão ou inclusão.

A Figura 2 apresenta o autômato utilizado para tratar a categoria dos artigos. Uma descrição detalhada da base para artigos se encontra na Tabela 1. No autômato da Figura 1, por exemplo, ao ser pesquisado o artigo "uns" será encontrado o "s" final[2], depois o "n" e depois um asterisco. Este símbolo indica a aceitação de qualquer caracter (ou qualquer conjunto de caracteres). No caso do artigo "uns", depois do "s" e do "n", a coincidência do asterisco ocorre com o caractere "u" que é, então, aceito. A próxima transição indica "–ns+m", o que significa a exclusão de "ns" e a inclusão de "m". Chega-se, assim, ao estado final, com a palavra "uns" sendo normalizada como "um".

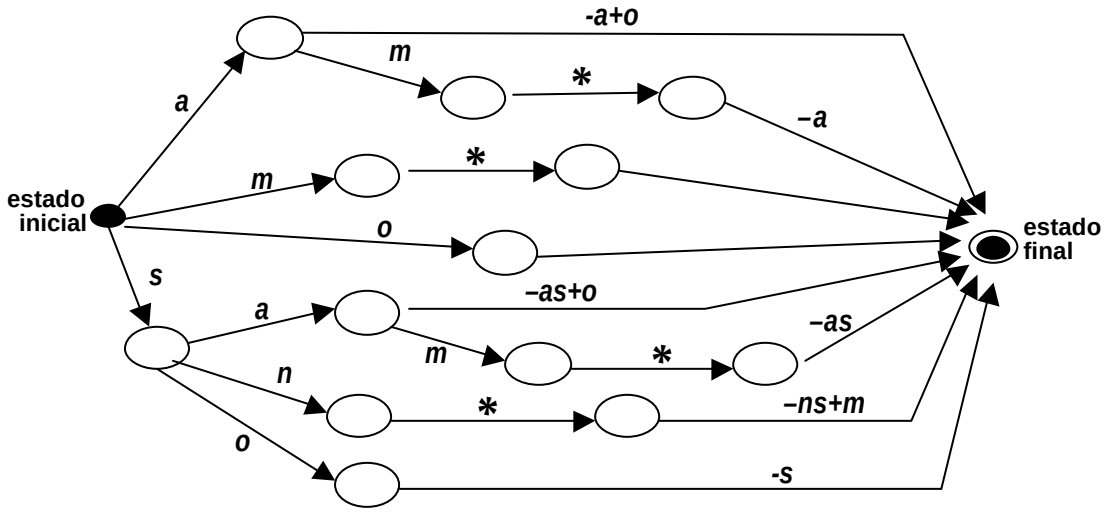

Figura 2. Autômato para normalização morfológica de artigos

3 Bases de sufixos

3.1 Conteúdo

O sistema utiliza bases de sufixos que são lidas para implementar os autômatos. Como já foi mencionado, existem seis bases específicas para as categorias morfológicas de palavras que apresentam variações: (1) adjetivos e verbos no particípio, (2) artigos, (3) numerais, (4) pronomes, (5) substantivos e (6) verbos (com exceção dos verbos no particípio). Todas elas apresentam, para cada sufixo armazenado, o mesmo conjunto de informações:

- ***Invertido***: caracteres do sufixo invertido (ou a terminação do item lexical a ser analisado, ou o item lexical inteiro, no caso de exceção);
- ***Exclusão***: ação de exclusão (representada por "–") de caracteres ao final do item lexical;

[2] A pesquisa é realizada do final para o início da palavra, para facilitar o reconhecimento do sufixo.

- ***Inclusão***: ação de inclusão (representada por "+") de caracteres ao final do item lexical; e
- ***Saída/Retorno***: ação de saída (representada por ">"), finalizando o processo, ou de retorno (representada por "<"), para continuar a normalização.

No campo ***Invertido*** é utilizado o caractere "=" para representar a repetição do caractere do sufixo anterior (ver exemplos a seguir). O asterisco, também usado neste campo, serve para representar qualquer conjunto de caracteres, como em "sies*>" que aceita, por exemplo, "seis" ou "dezesseis".

Em ***Saída/Retorno***, com "<", temos a indicação de que a palavra tratada ainda não está em sua forma normalizada. Os usos de "=" e "<" são detalhados adiante nas seções 3.3 e 3.4.

A base de sufixos completa para os artigos é apresentada na Tabela 1. Os registros desta base são utilizados para implementar o autômato da Figura 2.

Tabela 1. Base para artigos

Invertido	*Exclusão*	*Inclusão*	*Saída/Retorno*
a	-a	+o	**>**
=m*	-a		**>**
m*			**>**
o			**>**
sa	-as	+o	**>**
==m*	-as		**>**
=n*	-ns	+m	**>**
=o	-s		**>**

As demais bases de sufixos são sintetizadas na Tabela 2, com informações sobre a quantidade de registros e exemplos de normalização para cada base. No total, temos 1937 registros com sufixos invertidos e respectivas ações. A construção destas bases foi um trabalho exaustivo de pesquisa que, no entanto, não é dado ainda como finalizado.

Tabela 2. Síntese das bases de sufixos e alguns exemplos de normalização

bases		**Adjetivos**	**Numerais**	**Pronomes**	**Substantivos**	**Verbos**
registros		171	22	50	212	1474
exemplos	**palavra**	maníaca	primeira	minha	professora	cantará
	autômato	acaí*-a+o>	a*-a+o>	ahni*-inha+eu>	aro*-a>	ára*-á>
	saída	maníaco	primeiro	meu	professor	cantar
	palavra	limpinho	primeiros	algumas	carrinho	caibam
	autômato	ohni*-inho+o>	so*-s>	sam*-as>	ohni*-inho+o>	mabiac*-ibam+ber>
	saída	limpo	primeiro	algum	carro	caber
	palavra	felizes	seis	quaisquer	balões	chegou
	autômato	sezí*-es>	sies*>	reuqsiauq-isquer+lquer>	seõ*-ões+ão>	uo*-ou+ar>
	saída	feliz	seis	qualquer	balão	chegar

3.2 Busca de sufixos e exceções

Os registros das bases de sufixos nem sempre contêm apenas sufixos a serem pesquisados. Em alguns casos é preciso analisar outros elementos mórficos, como o radical da palavra. Quando é preciso analisar mais do que o sufixo, ocorre o que definimos como exceção. São consideradas exceções as palavras que não apresentam variações regulares, quando comparadas com o padrão da categoria. O tratamento das

exceções pode ser exemplificado através do trecho da base para substantivos apresentado na Tabela 3.

Tabela 3. Trecho da base para substantivos

Invertido	*Exclusão*	*Inclusão*	*Saída/Retorno*
...			
asac			>
==em*			>
==*	-a	+o	>
...			

No trecho apresentado na Tabela 3, temos alguns casos de sufixos terminados em "sa". A busca do sufixo invertido é processada de cima para baixo (do primeiro para o último registro) e a primeira coincidência faz com seja interrompido o processo. Assim, os registros da base de substantivos permitem que as exceções "casa", "mesa" e "sobremesa" (que não devem ser normalizados como "caso", "meso" e "sobremeso") permaneçam inalteradas. Outros substantivos terminados em "sa", como "dosa", serão normalizados através da substituição do "a" pelo "o". A repetição da terminação "sa" é indicada, nos dois registros seguintes (ver Tabela 3), pela cadeia "==", em "==em*" e em "==*".

Portanto, as exceções são tratadas junto à base da respectiva categoria morfológica da palavra tida como exceção. Dois exemplos, onde a variação irregular deve ser reconhecida por elementos mórficos diferentes do sufixo, são "reuqsiauq-isquer+lquer>", correspondente à palavra "quaisquer", e "mabiac*-ibam+ber>" que trata o verbo "caibam" (ver Tabela 2) .

3.3 Repetição de caracteres e tratamento de exceções

Nas bases de sufixos, o sinal de igual representa a repetição do caractere do registro anterior, na mesma posição contando da esquerda para a direita. Por exemplo, o campo ***Invertido*** no tratamento do sufixo "inho", para adjetivos, que deveria conter "ohni*-inho+o>", na verdade, contém

====*-inho+o>

porque aparece após outros registros com terminações semelhantes. Isto pode ser observado na Tabela 4, que apresenta um trecho da base para adjetivos com alguns casos terminados em "ho".

Tabela 4. Trecho da base para adjetivos

	Invertido	*Exclusão*	*Inclusão*	*Saída/Retorno*
	...			
1	ohl*			>
2	==nidnarg	-inho	+e	>
3	=====*	-inho	+o	>
4	====uq*	-quinho	+co	>
5	=====*	-uinho	+o	>
6	====zniur	-nzinho	+m	>
7	======ob	-nzinho	+m	>
8	=====*	-zinho		>
9	====*	-inho	+o	>
	...			

Nas oito linhas do campo ***Invertido*** (2 a 9) na Tabela 4, é utilizado o sinal "=" para representar a repetição de caracteres. Nestes registros (2 a 9), os dois últimos caracteres dos sufixos são "o" e "h". Estes caracteres ocorrem no registro 1 (onde estão explícitos) e nos registros 2 a 9, em virtude do "=". Nos registros 2 e 3 é repetido o final "ohnid" (na ordem correta, "dinho"). Nos registros 2 a 9 temos a repetição de "ohni" ("inho"), que somente surge com asterisco no registro 9, definindo a regra geral da normalização das palavras terminadas desta forma. Esta regra é usada no segundo exemplo para adjetivos, na Tabela 2.

Os sete registros, 2 a 8, apresentados na Tabela 4, são necessários para atender a exceções como as exemplificadas, na mesma ordem, a seguir:

"grandinho" (normalizado como "grande"),

"livrinho" (normalizado como "livro"),

"porquinho" (normalizado como "porco"),

"ruinzinho" (normalizado como "ruim"),

"bonzinho" (normalizado como "bom") e

"papelzinho" (normalizado como "papel").

Estas são exceções ao caso geral, que pode ser exemplificado pelo adjetivo "limpinho" (normalizado como "limpo") e que é atendido pelo último registro que aparece na Tabela 4.

3.4 Renovação da busca

Outra convenção encontrada nas bases de sufixos é a do retorno ao sistema de palavras parcialmente normalizadas. Esta ação, já mencionada, é indicada quando o campo ***Saída/Retorno*** contém "<" e não ">". Um exemplo é o do registro a seguir:

sahni*-as+o<

Este registro atende às palavras terminadas em "inhas". Note que "boazinhas" é uma exceção que não pode ser tratada por este registro. Mas, a palavra "limpinhas" é tratada aqui. Ela é parcialmente normalizada como "limpinho", ou seja, é passada do plural para o singular e do feminino para o masculino. Após, retorna para ter sua normalização completada pelo último registro da Tabela 4, sendo tratado então o diminutivo.

Outro caso pode ser exemplificado pela palavra "grandona" que, primeiramente, é reduzida para "grandão" (feminino para masculino) e, posteriormente, é normalizada como "grande" (tratando a forma aumentativa). A transformação direta, nestes casos, exigiria a repetição de trechos da base para tratar de palavras como "brincalhona" que não é normalizada como "brincalhe" (conforme seria de se esperar pela transformação válida para "grandona"). Estas decisões passam a ser tomadas em um único ponto de tratamento, associado somente ao sufixo "ão".

Do mesmo modo, o tratamento das palavras terminadas em "inhas" teriam que ter registros repetidos semelhantes àqueles (2 a 9) apresentados na Tabela 4. Estes, de acordo com o critério utilizado, ficam restritos a um único trecho da base de sufixos.

4 Avaliação

Um protótipo do sistema NILP pode ser encontrado em

http://www.inf.pucrs.br/~gonzalez/can.

A avaliação preliminar realizada utilizou um texto com 257 palavras etiquetadas morfologicamente. Este texto foi selecionado aleatoriamente a partir de uma coleção de documentos. Após a normalização executada pelo NILP, a saída do sistema foi analisada por três observadores humanos, que avaliaram a correção de cada item normalizado, de acordo com as normas gramaticais do Português. Neste contexto, a avaliação do protótipo obteve os seguintes resultados:

palavras desconhecidas = 1,2%
erros de normalização = 7,0%
normalizações corretas = 91,8%

As palavras desconhecidas foram mantidas na forma original. Nesta avaliação, isto pode ser considerado como acerto da normalização, já que os sufixos daquelas palavras não foram encontrados na base por dois motivos: (a) por conter erro de digitação (no texto original analisado) ou (b) por configurar nome próprio. As palavras digitadas de forma inválida teriam sido normalizadas corretamente, pois seus sufixos seriam encontrados nas bases utilizadas. Os nomes próprios teriam que ser realmente mantidos na forma original, como ocorreu. Portanto, recalculados os percentuais, podemos considerar uma precisão de 93%.

Esta é uma avaliação inicial realizada para justificar a continuação deste projeto. Outros testes, envolvendo o tratamento de *corpus* mais volumoso e a comparação com outros normalizadores, estão sendo preparados para o aprimoramento do sistema NILP.

5 Considerações finais

O sistema NILP se caracteriza por dispensar o uso de um léxico, porém exige que o texto de entrada seja etiquetado com as categorias morfológicas dos itens lexicais. Não temos notícia de outro normalizador (ou *stemmer*) que exija esta informação de entrada.

Os erros de normalização encontrados na avaliação do protótipo do sistema, e outros que devem surgir durante a fase atual de testes, são de fácil correção: na verdade remetem a uma alteração da base de sufixos correspondente, sem que haja necessidade de modificação no código do sistema. Esta é uma das vantagens que a estratégia adotada apresenta.

Duas outras vantagens, quando comparamos esta abordagem com a normalização através de *stemming*, podem ser encontradas no próprio conceito de redução à forma canônica e não em relação à estratégia adotada em si:

1) a manutenção da categoria morfológica original da palavra normalizada, que, em alguns casos, é uma necessidade ou mesmo uma exigência conforme a aplicação; e
2) a redução menos drástica (que acontece com a normalização à forma canônica), que elimina ambigüidades principalmente entre verbos e

substantivos quando, com *stemming*, é obtida a mesma forma reduzida, para palavras de categorias morfológicas originalmente diferentes.

Entretanto, para que o NILP seja transformado em um *stemmer*, são suficientes apenas alterações nas ações de exclusão e inclusão das bases de sufixos, sem que se altere a estratégia do sistema. É possível, inclusive, dispor de um conjunto de bases de sufixos para redução à forma canônica e outro para *stemming*.

Também é possível, com alterações nas ações de exclusão e inclusão da base para verbos, transformar o NILP em um sistema que recebe um verbo, como entrada, e devolve, como saída, o(s) substantivo(s) correspondente(s), configurando um processo de nominalização. Esta perspectiva está sendo estudada para que sejam desenvolvidos trabalhos futuros nesta direção, tendo o NILP como ponto de partida.

Deve ser salientado ainda que, sem alterar o algoritmo do sistema, é possível construir bases de sufixos para normalização morfológica de outros idiomas. Assim, apenas com a troca das bases, nosso normalizador deixaria de ser especializado na língua portuguesa e atenderia o idioma correspondente ao conjunto de bases utilizado.

Com a perspectiva da alteração de objetivo do NILP para *stemming*, para nominalização e para normalização morfológica de outros idiomas, outros testes podem ser planejados para validar nossa abordagem baseada em sufixos.

Por último, mencionamos que o presente trabalho, mesmo que não apresente ineditismo em seu resultado, foi desenvolvido por alunos de graduação em Ciência da Computação, servindo como uma introdução à problemática do processamento computacional da língua, e como motivação ao aprofundamento na área.

Referências bibliográficas

[Arampatzis2000] ARAMPATZIS, Avi. Linguistically-motivated Information Retrieval. Encyclopedia of Library and Information Science. Published by Marcel Dekker, Inc. - New York – Basel 2000, v. 69, p. 201-222

[Cegalla1998] CEGALLA, Domingos Paschoal. **Novíssima Gramática da Língua Portuguesa.** São Paulo, SP: Editora Nacional, 1998. 587p.

[Cohen1997] COHEN, D. I. A. **Introduction to Computer Theory**. New York: John Wiley & Sons, Inc., 1997

[Jurafsky2000] JURAFSKY, D.; MARTIN, J. **Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. New Jersey, USA: Prentice-Hall, 2000. 934 p.

[Lima1998] LIMA, C. H. R. **Gramática Normativa da Língua Portuguesa**. Rio de Janeiro: J. Olympio, 1998. 553p

[Orengo2001] ORENGO, V. M.; HUYCK, C. A Stemming Algorithm for the Portuguese Language. In: Eighth Symposium on String Processing and Information Retrieval (SPIRE 2001), Chile, 2001. P. 186-193.

[Sacconi1999] SACCONI, Luiz Antonio. **Nossa Gramática – Teoria e Prática.** São Paulo, SP: Atual Editora. 1999. 576p.