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

Lematização versus Stemming

J. L. De Lucca

Maria das Graças Volpe Nunes

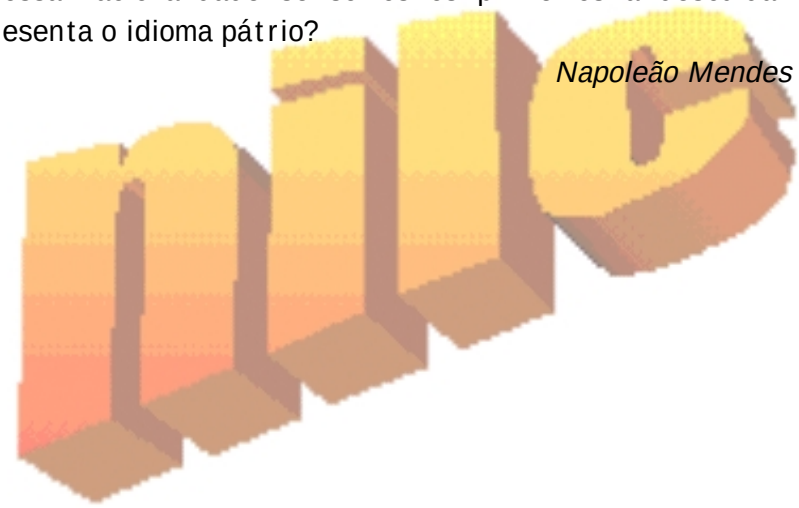
NILC-TR-02-22

Novembro 2002

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

A língua é a mais viva expressão da nacionalidade. Como havemos de querer que respeitem a nossa nacionalidade se somos os primeiros a descuidar daquilo que a exprime e representa o idioma pátrio?

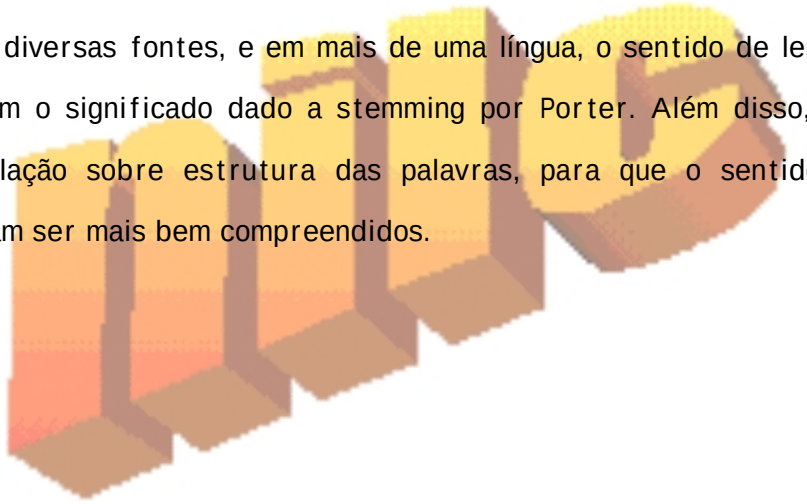
Napoleão Mendes de Almeida



Resumo

O objetivo deste Relatório Técnico é esclarecer as diferenças existentes entre lematização, forma canônica e stemming.

Para tanto, em diversas fontes, e em mais de uma língua, o sentido de lematização foi contrastado com o significado dado a stemming por Porter. Além disso, é feita uma breve recapitulação sobre estrutura das palavras, para que o sentido de lema e stemming possam ser mais bem compreendidos.



Índice

Lematização versus Stemming

1. Introdução

Lematização ou stemming? Esta é uma pergunta que muitos estudantes e professores se fazem quando perguntados sobre qual é a diferença entre stemming e lematização. Procuramos trazer definições de stemming, pelo próprio autor do algoritmo que traz o mesmo nome, PORTER (1980), assim como definições de lema em três línguas, de modo a deixar claras as diferenças existentes entre estas duas estruturas.

Embora se trate, aqui, de lematização, forma canônica e stemming, é trazido à luz de tal discussão um pouco de morfologia, para que o estudo da estrutura das palavras fique bastante claro para todos que se interessam pelo assunto.

2. Morfologia – Noções básicas sobre estrutura das palavras

Os elementos mórficos ou morfemas dividem-se em: raiz, radical ou tema, vogal temática, afixos e desinências.

Raiz – Em muitas gramáticas, o estudo das raízes é omitido, embora o termo esteja consignado na NGB (Nomenclatura Gramatical Brasileira). Raiz não é radical ou tema. Raiz é o elemento mórfico mais simples a que pode ser reduzida uma palavra. Obtém-se a raiz pela eliminação dos elementos secundários de formação.

Palavra	**raiz**
Abandonar	bann

Abandono	bann
Abandonadamente	bann
Abandonado	bann
Abdicar	dic
Abnegar	neg

Pode, não obstante, coincidir que o mesmo elemento venha a ser raiz e radical ou tema, ao mesmo tempo.

Palavra	**raiz, radical e tema**
Lavar	lav

Algumas palavras, com a evolução da língua, mantêm apenas uma letra da raiz original, como os exemplos abaixo.

Palavra	**Origem Latina**	**raiz**	**o que resta**
Feito	factu	fac	f
Malfeito	malefactu	fac	f
Feitura	factura	fac	f
Feitor	factore	fac	f
Benfeitor	benefactore	fac	f
Malfeitor	malefactore	fac	f

Radical – Radical é o elemento mórfico que fornece a significação da palavra. Por exemplo, nas palavras abaixo, o radical é comum entre todas elas: Pedr.

Palavra	**Radical**
Pedra	Pedr
Pedrinha	Pedr
Pedrada	Pedr

Abdicar	abdic
Abnegar	abneg

Tema

Tema é constituído pelo radical mais uma vogal temática à qual são acrescentadas as desinências casual para substantivos e adjetivos, e verbal para os verbos. Muitas vezes o tema coincide com o radical. Segundo COUTINHO (1954:171) "É comum entre nós não se distinguir tema de radical. Costumamos tomar as duas palavras como sinônimas".

Vogal temática

Vogal Temática é o elemento mórfico que se agrega ao radical de uma palavra para que ela possa receber outros morfemas. Divide-se em nominais e verbais.

Nominais – Referem-se a um substantivo ou adjetivo.

Palavra	**vogal temática**
Rosa	a
Poeta	a
Livro	o

Verbais – referem-se a um verbo.

Palavra	**vogal temática**
Cantar	a
Beber	e
Cair	i

Afixos

Afixos são elementos mórficos que se agregam a uma raiz ou radical a fim de mudar o sentido de uma palavra.

Os afixos subdividem-se em sufixos e prefixos.

Prefixos são antepostos ao radical.

Sufixos são pospostos ao radical.

Desinências

Desinência é o elemento mórfico que indica as flexões da palavra. Pode ser nominal ou verbal.

Nominal – Indica o gênero e o número dos nomes.

pat**a**
pat**os**

Verbal – Indica tempo, modo, número, pessoa e, também, as formas nominais do verbo.

Lembra**vas** desinência do pretérito imperfeito do indicativo
Lembr**ar** desinência do infinitivo
Lembr**ado** desinência do particípio

3. Lema, lematização e formas canônicas

No contexto puramente lexicográfico a palavra dicionarizada recebe a denominação de lema ou forma canônica. Assim, a representação gráfica das palavras recebe o nome de lemas ou formas canônicas, quando a palavra é representada pelo singular masculino para substantivos e adjetivos e infinitivo para verbos. A lematização é o ato de representar as palavras através do infinitivo dos verbos e masculino singular dos substantivos e adjetivos.

Segue, o significado ou conceito, conforme SAUSSURE (1955), de lema ou lematização, dado por diversos autores, em português, inglês e castelhano.

Para Gallison (1983), "Em Lexicologia: modo de agrupamento padrão das diversas variantes de um mesmo signo, com a finalidade de simplificar a apresentação e desse modo facilitar a consulta dos extratos lexicais em geral. Nos dicionários práticos, a lematização consiste em encontrar um item, isto é, uma forma gráfica representativa de todas as formas que uma unidade de significação lexicográfica (tradicionalmente palavra ou palavras compostas) pode tomar. É assim que o infinitivo é geralmente escolhido para simbolizar todas as formas do paradigma verbal (ex.: Ter por tenho, temos, terei, etc.); o masculino singular representa todas as formas do paradigma nominal e do paradigma adjetivo".

Este modo convencional de agrupamento das formas, indispensável em lexicografia, é por vezes incômodo em lexicologia descritiva (constituição de índice e de concordâncias e sobretudo em estatística lexical, porque registra numa única forma informações contáveis que não lhe pertencem propriamente e que seriam mais frutuosas se fossem manifestadas em cada uma das formas do paradigma. Quando se faz a extração de um corpus oral, em vez de se agruparem as ocorrências do verbo caminhar apenas sob a forma de infinito, poderia ser interessante, por exemplo, conhecer a distribuição comparada das formas do mais que perfeito simples e mais que perfeito composto na prática".

Assim como BIBER (1998:29) “The term «Lemma» is used to mean the base form of a word, disregarding grammatical changes such as tense and plurality. Thus, the frequency of the lemma DEAL – including *deal, deals, dealing,* and *dealt* – in the LOB Corpus is 290. (To distinguish between word forms and the lemma, we will use small capital letters to refer to the lemma DEAL.)”

Conforme The Oxford Dictionary and Thesaurus “Lemma is a heading indicating the subject or argument of a literary composition, a dictionary entry, etc.”

“Los lemas principales son los que encabezan los artículos. Corresponden siempre a la llamada forma canónica de una palabra. La forma canónica de los sustantivos y adjetivos es su forma de singular; la de los verbos es el infinitivo.” WERNER (2000).

“Junto con la identificación de las diferentes categorías morfológicas, las palabras que conforman un corpus pueden ser lematizadas (lemmatisation), es decir, pueden ser asignadas a su lema o forma canónica (la forma base que suele corresponderse con la entrada o voz en un diccionario). De este modo las diferentes formas flexivas del verbo trabajar, como por ejemplo trabajaba, trabajé, trabajaremos, etc. quedan resumidas en el lema trabajar y pueden incluirse todas en una sola búsqueda en el corpus.” PÉREZ(2002).

Sob o lema ou forma canônica é agrupado tanto o lema da palavra, propriamente dita, como as derivadas, as locuções, colocações etc., como na Fig. 1.

No parágrafo abaixo, por exemplo, há 63 palavras, mas somente 50 vocábulos – o processo de extrai-los é chamado de lematização -, o qual consiste em reunir todas as ocorrências da mesma palavra sob uma única forma, o lema, como acontece num dicionário, em vez de apresentá-las tal como aparecem nos textos, com variações no gênero, no número ou na grafia. DELUCCA (2001)

"Alguém definiu a linguagem como um conjunto de sons, grunhidos e outros barulhos, com os quais designamos metáforas. A humana é a única espécie que utiliza a linguagem dessa maneira. Os demais animais reconhecem as metáforas: um cachorro late para sua imagem no espelho porque acha que é outro cachorro. Mas não são capazes de traduzi-las em frases, poemas, contos, ensaios ou canções".

bacteria 35 **balk**

bacteria (*pronounced* bak-**tee**-ri-*a*) *noun plural* germs found in air, water, living and dead bodies, and especially decaying matter ▪ **bacterial** *adjective*

bad *adjective* **1** not good; wicked **2** not of a good standard: *bad workmanship/ bad at maths* **3** (often **bad for**) harmful: *smoking is bad for you* **4** of food: rotten, decaying **5** severe, serious: *bad dose of flu* **6** unwell ▪ **badly** *adverb* **1** not well **2** seriously **3** very much

(meaning 1) unpleasant, evil, naughty, disobedient; (meaning 2) poor, inferior, deficient, useless; (meaning 3) detrimental, adverse; (meaning 4) putrid, rancid

badge *noun* a mark or sign or brooch-like ornament giving some information about the wearer

badger *noun* a burrowing animal of the weasel family which comes out at night ▪ *verb* to pester or annoy

badminton *noun* a game resembling tennis, played with shuttlecocks

Although based on a 16th-century game, this was first played in its modern form in *Badminton* House in Avon

baffle *verb* **1** to be too difficult for; puzzle, confound **2** to prevent from being carried out; hinder ▪ **baffling** *adjective* (meaning 1)

bag *noun* **1** a holder or container, often of a soft material **2** a quantity of fish or game caught ▪ *verb* **1** to put in a bag **2** to secure possession of, claim: *bag a seat* **3** to kill (game) in a hunt

bag *verb* ▸ bags, bagging, bagged

bagatelle *noun* a board game, in which balls are struck into numbered holes

baggage *noun* luggage

bail[2] *verb*: **bail out** to bale out

bail[3] *noun*, *cricket* one of the crosspieces on the top of the wickets Probably comes from Old French *baillier* meaning 'to enclose', or from Latin *baculum* meaning 'stick'

Do not confuse with: **bale**

bailiff *noun* **1** an officer who works for a sheriff **2** a landowner's agent

bain-marie (*pronounced* ban-ma-**ree**) *noun* a double-boiler for cooking

Originally an alchemist's pot, named after the Biblical Mary, sister of Moses

bairn *noun*, *Scottish* a child

bait *noun* **1** food put on a hook to make fish bite, or in a trap to attract animals **2** something tempting or alluring ▪ *verb* **1** to put bait on a hook *etc* **2** to worry, annoy

baize (*pronounced* beiz) *noun* a coarse woollen cloth

bake *verb* **1** to cook in an oven **2** to dry or harden in the sun or in an oven

baker *noun* someone who bakes or sells bread *etc* ▪ **bakery** or **bakehouse** *noun* a place used for baking in

balaklava *noun* a knitted covering for the head and neck

balance *noun* **1** steadiness: *lost my balance and fell over* **2** the money needed to make the two sides of an account equal **3** a weighing machine ▪ *verb* **1** to be the same in weight **2** to make both sides of an account the same **3** to make or keep steady: *balanced it on her head*

balcony *noun* (*plural* **balconies**) **1** a platform built out from the wall of a building **2** an upper floor or gallery in a theatre *etc*

Fig. 1 Chambers Schoold Dictionary (sample page)

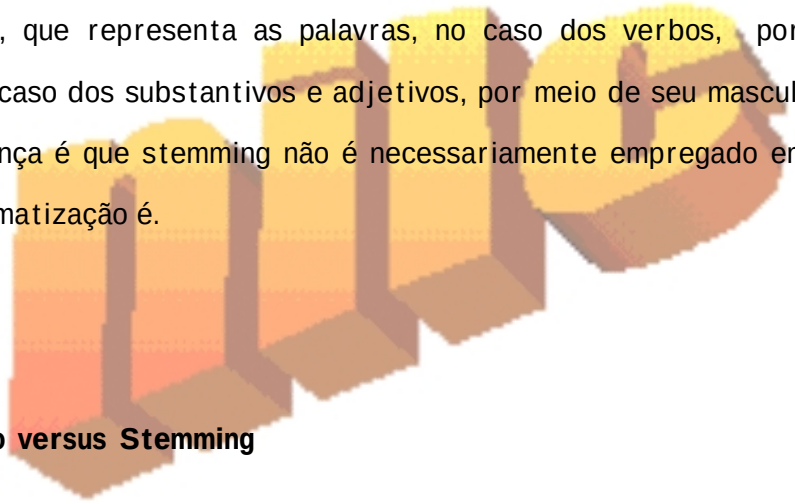
4. Stemmer

Em inglês, stem é o mesmo que radical ou tema. Raiz seria "root". Assim, para o The Oxford Dictionary and Thesaurus "**Stem** – 4. Gram. The root or main part of a noun, verb, etc. to which inflections are added; the part that appears unchanged throughout the cases and derivatives of a noun, persons of a tense, etc." Ainda de acordo com The Oxford Dictionary and Thesaurus, stemmer é a remoção de stems ou radicais.

The Cambridge International Dictionary of English define e exemplifica o que é stem:

"**STEM** CENTRAL PART The stem of a word is what is left when you take off the part which changes in order to show a different tense or a plural form etc.: From the stem 'sav-' you get 'saves', 'saved', 'saving' and 'saver'."

Para PORTER, "stemmer or 'Porter stemmer' is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems". O algoritmo de Porter é conhecido como "an algorithm for suffix stripping". Portanto, o algoritmo de Porter, na verdade, retira os sufixos das palavras, ao contrário da lematização, que representa as palavras, no caso dos verbos, por meio de seu infinitivo e, no caso dos substantivos e adjetivos, por meio de seu masculino singular. A segunda diferença é que stemming não é necessariamente empregado em lexicografia, ao passo que lematização é.



5. Lematização versus Stemming

Aqui podemos observar, no inglês, as diferenças gráficas entre stemming e lemmatizing, ou em outras palavras, remoção de sufixos, conforme PORTER, e lematização.

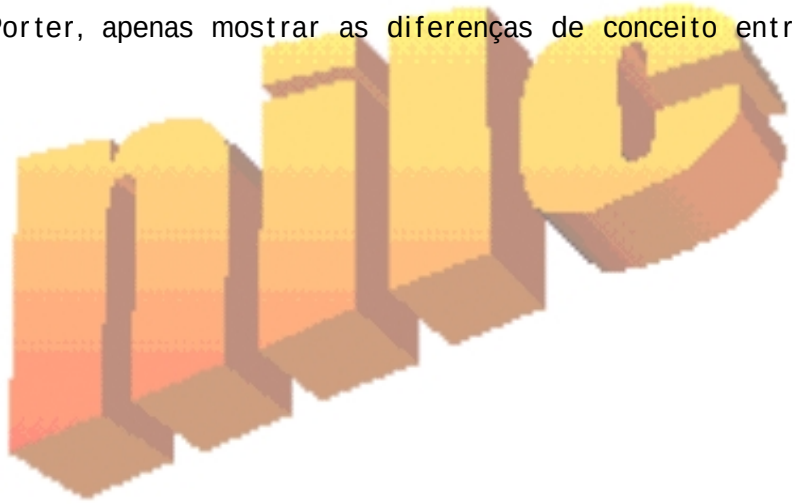
Seria interessante, aqui, colocar alguns exemplos em português também.

R. **O stem.c que eu possuo de Porter não compila.**

Palavra	**Lemmatizing**	**Stemming**
Agree	Agree	Agre
With	With	With
The	The	The
Policy	Policy	Polici
And	And	And
Dictionaries	Dictionary	Dictionari
Are	Be	Ar
More	More	More
Was	Be	Wa

Were	Be	Were
Speakers	Speaker	Speaker
Studying	Study	Studi
At	At	At
Separated	Separate	Separ
Saving	Save	Save
Saver	Save	Saver
Saved	Save	Save
Saves	Save	Save

Em alguns casos o algoritmo de Porter parece não reduzir a palavra ao seu radical, como em Saving, por exemplo, mas não é minha intenção discutir aqui o *modus operandi* algoritmo de Porter, apenas mostrar as diferenças de conceito entre stemming e lematização.



6. Conclusão

Lematização difere fundamentalmente de stemming. Enquanto lematização existe puramente no contexto lexicográfico, stemming não. Lematização é, pois, a representação da palavra através de seu masculino singular, adjetivos e substantivos e infinitivo (verbos), apenas no contexto da lexicologia. Stemming é a retirada de sufixos do radical, enquanto stem é o radical. Assim, as estruturas são distintas, embora eventualmente possam ser graficamente semelhantes.

7. Bibliografia

BIBER, D. et all. 1998. Corpus Linguistics – Investigating Language Structure and Use. Cambridge. Cambridge University Press.

CAMBRIDGE INTERNATIONAL DICTIONARY OF ENGLISH, THE. 1995. Cambridge. Cambridge University Press.

COUTINHO, I. L. 1954. Gramática Histórica. Rio De janeiro. Livraria Acadêmica.

DE CARVALHO, J. M. 1945. Dicionário Prático da Língua Nacional. Porto Alegre. Editora Globo.

DE LUCCA, J. L. 2001. Minidicionários da língua portuguesa: análise léxico-estatística, crítica e contrastiva das macro e microestruturas e sugestão de modelo. (Tese de Doutoramento, FFLCH/USP, São Paulo)

GALLISON, R. 1983. Dicionário de Didáctica das Línguas. Coimbra. Livraria Almedina

WERNER, R. Diccionarios Contrastivos del Español de América. Lehrstuhl für Angewandte Sprachwissenschaft (Romanistik) der Universität Augsburg http://www.answer.uni-augsburg.de/DCEA/decu/7-2.html. 2000.

MESQUITA, R.M. 1995. Gramática da Língua Portuguesa. São Paulo. Editora Saraiva.

OXFORD DICTIONARY AND THESAURUS,THE – Americam Edition. 1996. Oxford. Oxford University Press.

PÉREZ, C. 2002. Codificación (anotación y etiquetado) de los córpora http://elies.rediris.es/elies18/233.html

PORTER, M. F. 1980. An Algorithm for Suffix Stripping. **Progra**m. v. 14, n. 3, p. 130-137.

SAUSSURE, F. DE. 1955. Cours de linguistique générale. Lausanne, Payot. 5ª ed.