Capítulo 1. Radicalizadores versus analisadores morfológicos
Sobre a participação do Removedor de Sufixos da Língua Portuguesa nas Morfolimpíadas

Viviane Moreira Orengo e Diana Santos

1. Usos e história dos radicalizadores

A radicalização é um processo muito utilizado em Recuperação de Informações (RI). Sua aplicação baseia-se na suposição de que palavras com a mesma raiz são morfologicamente relacionadas. Por exemplo, se um usuário propõe a um sistema de RI a consulta "apresentação de artigos", é provável que este usuário também esteja interessado em documentos que contenham as expressões "como apresentar artigos" e "apresentando artigos".

O interesse em radicalizadores teve início nos anos 60. Um dos primeiros trabalhos nesta área foi o Algoritmo de Lovins (Lovins, 1968) para a radicalização do inglês. Nos anos seguintes vários outros algoritmos foram propostos. Dentre eles o mais utilizado é o Algoritmo de Porter (Porter, 1980), que baseia-se apenas num conjunto de regras, e apesar de ser um algoritmo simples, provou ser tão eficiente quanto sistemas mais complexos.

O processo de radicalização é executado por meio da remoção de afixos (prefixos e sufixos). Por exemplo, a forma *casas* pode ser reduzida à raiz "casa" através da remoção do sufixo *-s* indicativo de plural. Dois erros estão associados a este processo: a remoção indevida de caracteres que fazem parte da raiz e não do afixo (erro tipo 1), e a não remoção de caracteres que fazem parte do afixo (erro tipo 2).

Existem vários estudos avaliando o uso de radicalizadores para RI, e estes chegaram a conclusões contrastantes. Harman (1991) examinou o efeito de três algoritmos diferentes sobre três coleções de documentos e não observou melhoria no desempenho, já que o número de vezes em que a radicalização melhorou o resultado tende a igualar-se ao número de vezes em que o processo causou deterioração do resultado. Todavia, Krovetz (1993) observou melhorias de até 35% em algumas coleções de documentos. Finalmente, após extensa análise, Hull (1996) concluiu que "alguma forma de radicalização é quase sempre benéfica". Ele observou que o desempenho geral melhorou apenas de 1 a 3%, mas, ao analisar cada consulta individualmente, observou grandes melhorias.

Todos estes estudos foram realizados usando coleções em inglês, contudo é possível que línguas que têm mais formas variantes, tais como português, possam beneficiar mais do processo de radicalização. Porém, não temos conhecimento de trabalhos paralelos aos acima citados para o português.

2. Comparação com analisadores morfológicos

Para responder à questão de porquê um radicalizador participar nas Morfolimpíadas, é preciso identificar os pontos de semelhança e de diferenças entre os dois tipos de programas, e investigar se a diferença de tradições (radicalizar vs. analisar) tem origem no facto de os investigadores pertencerem a áreas diferentes (nomeadamente RI vs. PLN), ou se, pelo contrário, os objectivos são de facto diferentes.

Outra perspectiva, que abordaremos aqui, é ver se um analisador morfológico pode ser usado como radicalizador e, nesse caso, quais as vantagens ou desvantagens de tal opção.

E conversamente, investigaremos se a comparação com analisadores morfológicos poderá ser um bom (novo) paradigma para avaliar radicalizadores ou para ajudar no seu desenho.

Voltemos então a uma comparação entre os dois tipos de sistemas:

Em primeiro lugar, um analisador morfológico identifica um conjunto de traços pertinentes para a forma em questão (tempo e pessoa para formas verbais, pessoa e número para

formas nominais, etc.) e qual o lema (ou forma base, ou lexema) que permite representar a forma.

Um radicalizador, por seu lado, identifica uma raiz que idealmente deverá ser a mesma para formas relacionadas e que já está despida de desinências morfológicas. Geralmente, também pretende-se que a raiz vá mais longe do que uma simples (?) identificação do lema, de forma a juntar termos relacionados verbais, nominais e adjectivais, tal como "esquec", raiz comum às formas a que um analisador morfológico tradicional daria os lemas diferentes "esquecimento", "esquecer" e "esquecido"[1].

Se, por um lado, se compreende um analisador morfológico como um passo para uma análise sintáctica posterior em que o contexto irá desambiguar de que categoria gramatical se está em presença, um radicalizador, pelo contrário, é um passo para associar termos aparentemente diferentes num mesmo índice, para facilitar e aliviar os mecanismos de recolha de informação.

Vemos portanto que, mesmo se olharmos apenas para a detecção do lema (um subconjunto dos requisitos de um analisador morfológico), enquanto um analisador morfológico terá como objectivo discriminar diferentes lemas com base na categoria gramatical, um radicalizador terá como requisito obter radicais diferentes apenas se estivermos em presença de conceitos radicalmente diferentes.

Daí que o passo inicial para comparar a saída de um analisador morfológico com a de um radicalizador é a de tentar obter uma raiz a partir dos lemas. Na secção 5, usando os resultados obtidos nas Morfolimpíadas, tentamos fazer exactamente isso.

Outra diferença, já implicitamente descrita acima, mas que é importante reforçar, é que normalmente um analisador morfológico não se preocupa com a semântica, produzindo pois o mesmo resultado para *canto* ou para *aterrar*, independentemente dos vários sentidos de *canto* (p. ex, em *o canto mavioso dos rouxinóis*, *o canto sexto dos Lusíadas* ou *o canto mais afastado da sala*) e de *aterrar* (inspirar terror ou pousar em terra). Já um radicalizador (que nestes dois casos também não produziria diferença) é geralmente avaliado exactamente pela capacidade de agrupar correctamente termos relacionados ou (negativamente) pelo número de termos que agrupou incorrectamente[2].

Finalmente, e dependendo do funcionamento interno de um sistema de procura de informação, a eficiência deste (e da radicalização envolvida) pode ser relevante, levando à escolha de sistemas menos pesados. Devido ao aumento exponencial do poder de cálculo e de memória nos sistemas computacionais (ver a lei de Moore), não trataremos deste aspecto aqui, por nos parecer ter já sido ultrapassado.

3. Avaliação de radicalizadores

O método tradicional de avaliação de radicalizadores empregado em RI consiste em aplicar o radicalizador a uma coleção de teste, calcular precisão e cobertura ("recall") em relação a um dado conjunto de procuras ("queries"), e avaliar o efeito do radicalizador sobre estas medidas. Um sistema é considerado melhor do que outro se produzir maior precisão e cobertura. O problema deste método é que não permite a identificação dos erros cometidos pelo radicalizador, portanto não ajuda o programador na otimização de seus algoritmos. Além disso, como já indicado na secção 1, os resultados não são indiscutíveis. Para o inglês, tanto Harmam (1991) como Woods *et al.* (2001) declaram que o uso de radicalizadores em alguns casos melhora, mas noutros degrada o desempenho, e que é preciso uma análise mais fina e estratégias diferentes (por exemplo, em alguns casos, não radicalizar).

Outra possibilidade seria manualmente atribuir a raiz correta a uma lista de palavras e compará-las com o resultado produzido pelo radicalizador, como feito em Chaves (2003). Este método apresenta, no entanto, alguns problemas. Primeiramente, o processo de manualmente

[1] De facto, é interessante salientar que alguns analisadores morfológicos mais inovadores, como o Jspell (Almeida & Simões, neste volume) propõem o lema "esquecer" para as três formas.

[2] Por isso, convém salientar que o processo de radicalização está intimamente relacionado com as características da língua, que lhe colocam limites de desempenho inultrapassáveis sem auxílio do contexto.

atribuir a raiz correta é bastante trabalhoso. Além disso, nem sempre é óbvio qual deve ser a raiz correta para algumas formas, ou seja, o mais importante é que a raiz seja uma ajuda para o agrupamento, esse sim capaz de ser avaliável independentemente.

Silva (2001), comparando dois radicalizadores diferentes, mediu o grau de agrupamento sobre uma lista de 7458 palavras distintas, definido como o número médio de formas por radical, tendo obtido os valores bem distintos de 1,15 e 14,48 (ver também Silva & Oliveira, 2003). Silva (2001:95) não deixa contudo de frisar que este é apenas um aspecto parcial, visto que a coerência dos agrupamentos é fulcral. É preciso, contudo, notar que Silva aceita mais de um radical por forma (ele de facto fala em lematização), donde não é evidente que nos encontremos em presença do mesmo tipo de sistema discutido neste artigo.

Um terceiro método de avaliação foi proposto por Paice (1994). Este método requer uma amostra de palavras separadas em grupos. Cada grupo deve conter palavras semanticamente relacionadas entre si. Após o processo de radicalização o algoritmo ideal deveria:

(i) reduzir todas as palavras de um grupo à mesma raiz;

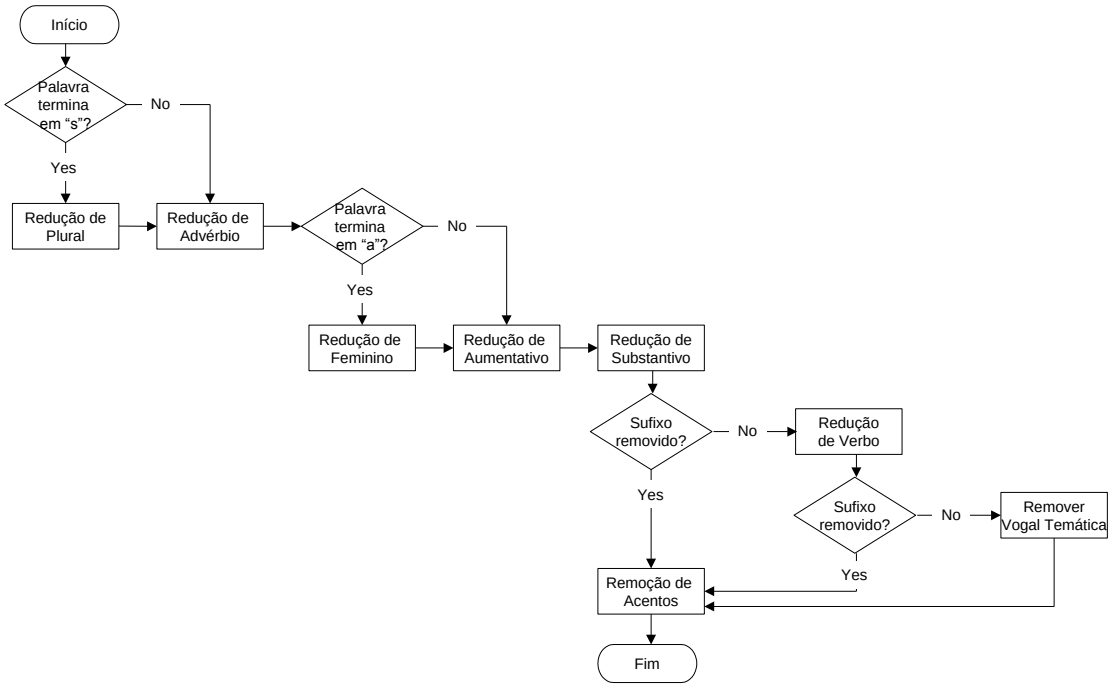
(ii) atribuir uma raiz individual a cada grupo, ou seja, a raiz atribuída a um grupo não deve coincidir com a raiz atribuída a outros grupos.

Se (i) e/ou (ii) não ocorrerem, significa que o radicalizador cometeu erros. O método fornece medidas para quantificar o desempenho do radicalizador quanto aos erros dos tipos 1 e 2 (ver seção 1). Desta maneira um algoritmo pode ser considerado melhor do que outro se cometer menos erros de ambos os tipos.

Assim como o método anterior, a avaliação de Paice requer a intervenção humana sendo portanto muito trabalhosa. Sua maior vantagem é fornecer meios para que o programador identifique os erros cometidos pelo radicalizador, possibilitando assim a otimização do algoritmo.

4. O Removedor de Sufixos da Língua Portuguesa

Esta seção descreve o Removedor de Sufixos da Língua Portuguesa (RSLP), inicialmente proposto em Orengo & Huyck (2001)[3]. O algoritmo foi implementado em C e é composto por 8 estágios que devem ser executados seqüencialmente. A Figura 1-1 ilustra a seqüência destes



[3] Para a lista completa das regras, bem como o algoritmo em C, refira-se ao endereço eletrônico http://www.cs.mdx.ac.uk/research/PhDArea/RSLP

estágios. Cada estágio é composto por uma série de regras e em cada estágio apenas uma regra é aplicada. O algoritmo baseia-se apenas nestas regras, não fazendo consultas a dicionários.

Figura 1-1: Estágios do Removedor de Sufixos

Estágio 1 - Redução de Plurais: Com raras exceções, os plurais em Português terminam em *-s*. Entretanto, nem todas as palavras terminadas em *-s* são plurais (ex. *lápis*). Este estágio consiste basicamente em remover o *-s* final das palavras que não estão presentes nas listas de exceções. Algumas vezes outras modificações são necessárias, por exemplo, palavras terminadas em *-ns* devem ser este sufixo substituído por *-m* (ex. bons ? bom).

Estágio 2 - Redução de Advérbios: Este é o estágio mais curto de todos, já que há apenas um sufixo que denota advérbio *-mente*. Assim como no estágio anterior, nem todas as palavras terminadas em *-mente* são advérbios, por isso uma lista de exceção faz-se necessária.

Estágio 3 - Redução de Femininos: Este estágio consiste em transformar as formas femininas para a forma masculina correspondente. Apenas palavras terminadas em *-a* passarão por este estágio.

Estágio 4 - Redução de Aumentativos e Diminutivos: Conforme Cunha & Cintra (1985), existem 38 sufixos que denotam aumentativos e diminutivos, contudo vários estão obsoletos. Assim, este estágio remove apenas os casos mais comuns que ainda se encontram em uso.

Estágio 5 - Redução de Substantivos: Este estágio remove 61 sufixos comuns a adjetivos e substantivos. Palavras que têm sufixo removido por este estágio não passarão pelos estágios 6 e 7, indo diretamente para o estágio 8.

Estágio 6 - Redução de Verbos: De todas as categorias, os verbos são os que apresentam o maior número de formas variantes. Os verbos regulares podem apresentar mais de 50 formas diferentes. Este estágio reduz os verbos a sua raiz, já que seria impossível reduzi-los ao seu infinitivo sem consultar um dicionário.

Estágio 7 - Remoção de Vogal: Este estágio consiste em remover a última vogal (“a”, “e”, ou “o”) de palavras que não foram tratadas nos estágios 5 e 6. Por exemplo, a palavra *menino* não teria sofrido nenhuma modificação nos passos anteriores. Neste estágio ela perderia o *-o* final para poder combinar-se com outras formas variantes como *menina* e *meninice* que teriam sido reduzidas para “menin”.

Estágio 8 - Remoção de Acentos: A remoção de acentos é necessária porque existem casos em que algumas formas variantes da mesma palavra são acentuadas e outras não, como *psicólogo* e *psicologia*. É importante que este passo seja executado no final do processo de radicalização e não no princípio porque a presença de acentos é significante para algumas regras, por exemplo óis?ol que reduz *sóis* para “sol”.

As regras de radicalização foram criadas com base nos sufixos mais comuns apresentados em Cunha & Cintra (1985) e Macambira (1999). A versão atual do algoritmo conta com 199 regras.

Cada regra estabelece:

- O sufixo a ser removido;
- O tamanho da menor raiz, para evitar de remover um sufixo e deixar uma raiz pequena demais.
- Um sufixo substituto, que por vezes faz-se necessário;
- Uma lista de exceções. Existem exceções para quase todas as regras estabelecidas, desta maneira, listas de exceções fizeram-se necessárias para evitar erros do tipo 1.

A seguir temos um exemplo de regra de radicalização:

```
"inho", 3, "", {"caminho", "carinho", "cominho", "golfinho", "padrinho",
"sobrinho", "vizinho"}
```

Onde *-inho* é o sufixo que denota diminutivo; 3 é o menor tamanho permitido a uma raiz (o que previne que palavras como *vinho* sejam reduzidas); e as palavras entre chaves são as exceções a esta regra, ou seja, palavras que terminam em *-inho*, mas onde este não é sufixo e sim parte da raiz.

Uma importante observação é que as raízes não precisam ter significado lingüístico, já que elas são utilizadas para indexar uma coleção de documentos e não serão mostradas ao usuário. Contudo, elas precisam captar o significado da palavra sem perder muitos detalhes.

Para que o RSLP participasse na avaliação conjunta, foi apenas preciso, para o caso dos átomos que o RSLP não contemplava, tais como sinais de pontuação e caracteres raros, ecoá-los na saída, de forma a poder ter um resultado para cada unidade, ou seja, obter o mesmo formato (um formato comparável ao) das Morfolimpíadas.

Isso transformou a saída de pontuação, não analisada pelo sistema, idêntica aos casos em que o radical era o mesmo, o que nos parece conceptualmente correcto.

5. Os resultados nas Morfolimpíadas

Como indicado na secção 2, seria em princípio possível obter um pseudo-radicalizador a partir de um analisador morfológico, se apenas nos concentrássemos no(s) lema(s) devolvido(s) por estes.

Foi assim criado pela organização mais um conjunto de resultados, sob o título "Os sistemas como radicalizadores", em que, junto com o RSLP, se encontra a saída dos vários analisadores morfológicos, seguindo o seguinte algoritmo:

Se só houvesse uma análise para a forma (por exemplo, *castos*), ou se todas as análises correspondessem ao mesmo lema (por exemplo, *castiga*), devolvia-se o lema. Se houvesse intersecção entre os vários lemas diferentes (por exemplo, *casta*, essa era a raiz, senão devolvia-se a forma não analisada (por exemplo, *fora*, seguida pela marca NC). Ou seja, e exemplificando com o seguinte analisador morfológico abstracto, a saída como radicalizador das seguintes formas seria a apresentada na Tabela 1-1.

Tabela 1-1: Exemplos de conversão de um analisador morfológico num radicalizador

Forma	Resultado Analisador Morfológico	Resultado Radicalizador
cantiga	cantiga÷SUB÷cantiga÷.÷S÷.÷F÷.÷.÷.÷.÷.	cantiga => cantiga
casta	casta÷ADJ÷casto÷.÷S÷.÷F÷.÷.÷.÷.÷. casta÷SUB÷casta÷.÷S÷.÷F÷.÷.÷.÷.÷	casta => cast
castiga	castiga÷V÷castigar÷IMP÷S÷2÷.÷.÷.÷.÷.÷. castiga÷V÷castigar÷PR_I÷S÷3÷.÷.÷.÷.÷.÷.	castiga => castigar
castos	castos÷ADJ÷casto÷.÷P÷.÷M÷.÷.÷.÷.÷.	castos => casto
fora	fora÷ADV÷fora÷.÷.÷.÷.÷.÷.÷.÷. fora÷INTERJ÷fora÷.÷.÷.÷.÷.÷.÷.÷. fora÷V÷ir÷PSM_I÷S÷1÷.÷.÷.÷.÷.÷. fora÷V÷ir÷PSM_I÷S÷3÷.÷.÷.÷.÷.÷. fora÷V÷ser÷PSM_I÷S÷1÷.÷.÷.÷.÷.÷. fora÷V÷ser÷PSM_I÷S÷3÷.÷.÷.÷.÷.÷.	fora => fora NC

5.1 Contabilização do resultado do sistema

Para poder comparar os sistemas, e indicar os casos em que eles produzem algo não trivial, contámos a) o número de casos em que o radical é diferente da forma (e que, portanto, a informação produzida é potencialmente relevante), assim como b) o número de casos em que não foi produzida qualquer intersecção (e que, portanto, o conhecimento morfológico indica que são ou deveriam ser ambíguos entre vários radicais diferentes).

Também contámos ou tentámos medir a ambiguidade morfológica no que diz respeito ao lema, ou seja, c) em quantos casos (segundo cada analisador) uma palavra tem um lema único e d) em quantos casos esse lema depende das suas características morfológicas.

Neste caso, era francamente fácil a um observador independente distinguir a saída de um verdadeiro radicalizador da saída dos analisadores morfológicos virados radicalizadores, não só pela inexistência de marcações "NC" na saída como pela inexistência de radicais "perfeitos" (ou seja, lemas). Assim podemos indicar que o sistema H corresponde ao RSLP, e comparar com os outros, que mantemos naturalmente anonimizados.

Para neste caso poder comparar com mais fiabilidade o radicalizador, usámos a atomização produzida pela organização (sobre o ficheiro `uul.txt`, uma unidade por linha), mas indicamos também os valores para o número de formas distintas, descontando a pontuação[4].

Os resultados dos vários sistemas encontram-se na Tabela 1-2 e na Tabela 1-3[5]. Como já referido, o sistema H é o RSLP, os outros são analisadores morfológicos.

Tabela 1-2: Resultados globais sobre a radicalização dos sistemas: Formas

Sistema	Formas	Irradicalizáveis		Iguais		Diferentes	
A	64704	1367	2,11%	41737	64,50%	21600	33,38%
B	67609	584	0,86%	39223	58,01%	27802	41,12%
C	65643	1417	2,16%	43967	66,98%	20259	30,86%
D	66943	1330	1,99%	44569	66,58%	21044	31,44%
E	65035	2720	4,18%	42148	64,81%	20167	31,01%
F	66478	1036	1,56%	45659	68,68%	19783	29,76%
H	68755	0	0%	42	0,06%	68713	99,94%

Tabela 1-3: Resultados globais sobre a radicalização dos sistemas: Tipos

Sistema	Tipos	Irradicalizáveis		Iguais		Diferentes	
A	17028	176	1,03%	7522	44,17%	9330	54,79%
B	15799	28	0,18%	4711	29,82%	11060	70,00%
C	15628	133	0,85%	6627	42,40%	8868	56,74%
D	13674	124	0,91%	5295	38,72%	8255	60,37%
E	14341	991	6,91%	5217	36,38%	8133	56,71%
F	15963	141	0,88%	7108	44,53%	8714	54,59%
H	16632	0	0%	8	0,05%	16624	99,95%

Conforme seria de esperar, enquanto o RSLP produz um radical menor (e portanto diferente) da palavra na quase totalidade dos casos (99,94%), a situação no caso dos analisadores morfológicos é bem distinta, produzindo estes um radical diferente da forma apenas de 29,76% a 41,12% dos casos.

Por outro lado, encontramos de 4,18% a 0,86% de casos irradicalizáveis, no sentido de que não há uma cadeia de caracteres (o desejado radical) que cubra as várias acepções morfológicas da palavra. Exemplos são: *fosse* (cujo lema pode ser "ir" ou "ser"), *vá* (cujo lema pode ser "ir" ou "vá", considerado como interjeição), *piores* (cujo lema pode ser "piorar" ou "mau"), *vão* (cujo lema pode ser "vão" ou "ir"), *retratei* (cujo lema pode ser, para sistemas que tratam a derivação, "tratar" ou "retratar"), assim como palavras em maiúscula consideradas

[4] Note-se que os resultados dos outros sistemas foram executados sobre o ficheiro `ts.txt` (texto seguido), ou seja, seguiu-se a atomização própria de cada sistema. No caso do RSLP, usámos a atomização da organização, visto que os caracteres « e », próprios apenas da variante de Portugal, que foram abundante e consistentemente usados nos textos das Morfolimpíadas para codificar as aspas, eram acoplados pelo RSLP à palavra, produzindo um maior número de formas e desse modo dificultando a comparação com os outros sistemas:

```
«existe => exist
estrangeiros» => estrang
```

[5] Estes valores foram calculados desprezando os clíticos, devido à grande variabilidade dos lemas destes, e ao facto de nos parecer natural que fosse o verbo só a decidir o agrupamento, e portanto o radical.

pelos sistemas como podendo ter um lema minúsculo (o nome próprio *Marco* poder ter lema "marco" ou "Marco")[6].

Esta primeira comparação não diz contudo muito sobre os méritos (ou defeitos) de um radicalizador, sobretudo se atentarmos no facto de as formas irradicalizáveis acima citadas não parecerem encontrar-se entre os termos típicos de uma procura. Além disso, se as formas irradicalizáveis sofrerem uma radicalização forçada, perdem por definição uma ou mais acepções, o que será mais ou menos problemático conforme a informação de que se está à procura.

Uma maneira mais esclarecedora será comparar os sistemas na sua capacidade de agrupamento, definindo capacidade de agrupamento (em função do corpus) como o número de grupos com mais de duas palavras encontrados pelos radicalizadores.

5.2 Comparação em termos dos grupos obtidos no corpus das Morfolimpíadas

Para as formas únicas (tipos), apresentamos na Tabela 1-4 as seguintes estatísticas: número total de palavras, número de palavras sós, número de palavras agrupadas, número de grupos distintos, capacidade de agrupamento (número de grupos com duas ou mais palavras), factor de compactação (razão do número de palavras pela soma de todas as palavras, valendo as sós 1, e as pertencentes a um grupo de *n* elementos $^1/_n$).

De notar que estes valores são independentes da forma do radical. Eles medem, sim, a capacidade de agrupamento seja qual for o radical.

Tabela 1-4: Capacidade de agrupamento dos vários sistemas

Sistema	Palavras total	Número de grupos ≥ 2	Número de grupos total	Número de palavras nos grupos	Factor de compactação
A	17485	2752	11992	8244	1,46
B	17053	2930	10082	8978	1,69
C	16981	2414	10396	8021	1,62
D	15704	2302	9680	6596	1,62
E	16368	2243	10420	6464	1,57
F	17324	2566	10498	8383	1,65
H	17132	3133	8713	11546	1,97

Na Tabela 1-4, provavelmente o que chama mais a atenção é o enorme número de grupos com uma palavra só. Tal deve-se a razões diversas: a quantidade de nomes próprios e de números é considerável (aprox. 2500); depois, conforme já a lei de Zipf nos deixaria adivinhar, muitas palavras aparecem apenas uma vez (e/ou numa forma), donde para o caso deste corpus (que é muito pequeno) a radicalização apenas interessaria a cerca de 17% das unidades. Seja como for, esta observação é válida para todos os sistemas, donde a comparação é justa.

Olhando para os grupos obtidos podem-se imediatamente distinguir dois tipos de funcionamento. Escolhendo arbitrariamente o (analisador morfológico como) radicalizador C, ilustramos o seu funcionamento na Tabela 1-5, observando vários casos de radicais diferentes que correspondem (ou podem corresponder) a sentidos diferentes, veja-se (1-4), (8), (12), (14). Por outro lado, é frequente a situação em que mais do que um radical poderia com vantagem ser amalgamado, como o ilustram os casos (5-7),(9), (10-11), (13), (15) e possivelmente (16).

Tabela 1-5: Alguns conjuntos de grupos obtidos pelo "radicalizador" C. Os números são para referência no texto

	Radical		Formas
1	acha	2	*acha achados*
	achar	16	*achá-la achá-lo acham achamos Achámos achará achar acharia achava Achavam achava-se Achei acho Acho achou achou-se*
2	Acord	3	*acordes acordo desacordo*

[6] De facto, poder-se-ia considerar pertinente fazer uma transformação prévia de todas as palavras para a sua versão minúscula, como é, aliás, o caso do funcionamento do RSLP. Contudo, assim cometer-se-ia o erro de juntar palavras sem qualquer relação, distinguidas aliás por uma grafia diferente. Na prática, ambas as soluções têm vantagens e inconvenientes.

	acordar	6	*acorda acordado acordara acordar acordei acordou*
3	Acto	2	*acto actos*
	actor	5	*actora actor actores ator atores*
	actriz	2	*actriz atriz*
4	actua	4	*actuação actuais actuantes atuais*
	actual	2	*actual atual*
	actuar	2	*actua actuarão*
5	agrad	2	*agradável agrado*
	agradar	5	*agradam agradasse agradava agradáveis desagradáveis*
6	ajust	2	*ajuste Ajuste*
	ajustar	6	*ajustada ajustado ajustamento ajustar Ajustar ajustáveis*
7	alega	2	*alegada alegados*
	alegar	3	*alega alegando alegarem*
8	alto	4	*altas altíssimo alto altos*
	altura	2	*altura alturas*
9	am	3	*ama AMA amo*
	ama	2	*amadores amante*
	amar	7	*amada amado amamos amara amar amava amável*
10	amig	4	*amiga amigas amigável amigo*
	amigo	3	*amicíssimo amigos amiguinho*
	amistoso	2	*amistosa amistosos*
11	ambient	2	*ambiente ambientes*
	ambiental	2	*ambiental ambientalistas*
12	animal	2	*animal animalidade*
	animar	4	*animação animada animadoras animaria*
13	anunciar	5	*anunciado anunciam anunciaram anunciar anunciou*
	anúncio	2	*anúncio anúncios*
14	aparar	2	*aparada apare*
	aparecer	12	*apareça aparece Aparecem aparecendo apareceram aparecerão aparecer apareceria apareceu aparecia apareciam aparecimento*
	aparência	2	*aparência aparências*
15	aplica	2	*aplicação aplicada*
	aplicar	6	*aplicações aplicado Aplicando aplicar aplica-se aplicava-se*
16	apo	2	*aposta apostas*
	apoderar	2	*apodera-se apoderava-se*
	apoi	2	*apoio APOIO*
	apoiar	5	*apoiada apoiar apoiarem Apoiemos apoiou*

Pelo contrário, um radicalizador de raiz (passe o pleonasmo), ilustrado na Tabela 1-6, peca em geral no sentido contrário, ou seja, produz radicais demasiado abrangentes, como é o caso de (6-8), (10-11) e (13), ou, pelo contrário, funciona amalgamando um conjunto certo (1-5), (9), (12) com um radical mais "radical" do que o produzido pelos analisadores morfológicos.

Tabela 1-6: Alguns grupos obtidos pelo RSLP

	Radical	Formas	
1	abert	7	*aberta Aberta abertamente abertas aberto abertos abertura*
2	abol	2	*abolia abolicionista*
3	abomin	2	*abomina abomináveis*
4	abord	3	*abordagem abordam abordando*
5	aborrec	5	*aborrece aborrecem aborrecida aborrecido aborrecimento*
6	ach	13	*acha achados acham achamos achará achar acharia achava Achavam Achei acho Acho achou*
7	acid	4	*acidente Acidente acidentes ácido ácidos*
8	acord	10	*acorda acordado acórdão acordara acordar acordeão acordei acordes acordo acordou*
9	acresc	4	*acresce Acresce acrescidas acréscimos*
10	act	7	*acto actora actor actores actos actuais actual*
11	adiant	4	*adianta adiantadas adiante adiantou*

12	admir	6	*admiração admirado admirador admirava admiráveis admirável*
13	aflu	2	*afluência afluente*

Se, em vez de uma análise impressionista como a acima, olharmos antes para os grupos com mais membros, uma conclusão torna-se evidente: é precisamente nos casos dos verbos irregulares que os analisadores morfológicos são excelentes, senão vejam-se os grupos TER, IR, DAR e FAZER na Tabela 1-7.

Tabela 1-7: Grupos com mais exemplares e sua contrapartida no RSLP

Grupo	Número de exemplares pelos AM	Número de exemplares pelo RSLP
TER	39, 38, 37, 28	8
IR	28,26, 25, 22, 14	4
DAR	38, 31, 32, 30, 29	6
FAZER	38, 37, 35, 29	far 5 faz 13 faç 3
PODER	26, 24, 23, 20	pod 22 poder 4 poss 6 pud 5
SER	33, 31, 30, 24	14

Outros grupos com vários exemplares são: VER (27), DIZER, ESTAR, DEIXAR (26), FICAR, ENCONTRAR (25), CONHECER (24), QUERER, CHEGAR, DEVER, SABER (22), HAVER, VIR (21) e SENTIR (20).

Por outro lado, os maiores sucessos de um radicalizador genuíno, apresentados na Tabela 1-8, são casos de verbos regulares e casos em que nomes, verbos e adjectivos partilham suficiente forma e sentido para valer a pena serem agrupados:

Tabela 1-8: Tamanho dos grupos obtidos pelo radicalizador

pass	29
cas	23
cont	20
continu	22
dev	23
viv	23
cheg	20
deix	17
lev	20
olh	20
pod	21

Mas se *contar*, *contas* e *contagem*, por um lado, e *continuar*, *continuação*, *continuamente* e *continuidade*, por outro, parecem suficientemente relacionados para serem justamente agrupados, já o mesmo pode ser questionável para *olhos* e *olhar* ou para *passar* e *passado*. Isto porque, embora haja uma forte relação entre estas palavras, não é óbvio que ao procurar informação sobre uma se esteja geralmente interessado na outra.

Em Santos (1996) tentou-se uma primeira classificação das relações de homografia em termos de relação semântica com base num pequeno corpus, que mede evidentemente um caso limite que nenhum radicalizador pode ultrapassar. Os números indicam que aproximadamente 20% das palavras homógrafas não estão semanticamente relacionadas, e que em apenas 33% dos casos estão relacionadas por derivação (directa ou contendo uma origem comum). A serem representativos estes valores, isto significa que, nos casos de homografia transcategorial (ajuizados em 18,7%, não contando com nome vs. adjectivo, em Medeiros *et al.*, 1993), o agrupamento de um terço (ou seja, 6%) das palavras dará sempre resultados errados visto que a palavra contém em si grupos distintos. Isto é uma das razões porque mesmo o método de Paice, porque assume um grupo por palavra, não permite avaliar toda a problemática da radicalização[7].

[7] Por exemplo, a palavra *juro* deve ser junta a *juros* ou a *jurar*? Isto depende do interesse do utilizador, mas também da frequência relativa das duas interpretações nos textos sobre os quais se procura.

5.3 Comparação em termos do tamanho dos grupos

Em termos de tamanho dos grupos, a Figura 1-2 mostra a sua variação para cada sistema. No eixo dos XX temos o tamanho do grupo, no eixo dos YY o número de grupos distintos com esse tamanho.

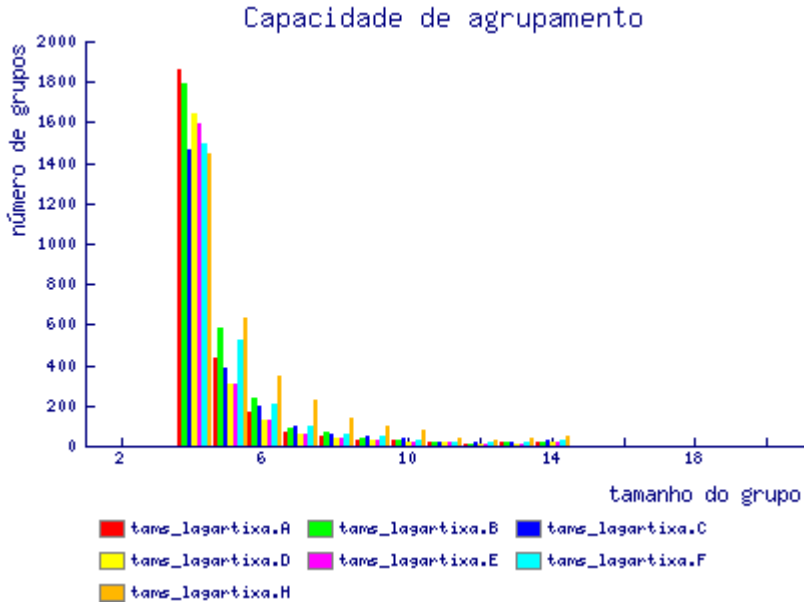


Figura 1-2: Número de grupos com mais de um elemento obtidos por cada sistema, por tamanho

5.4 Comparação usando o método de Paice

Finalmente, usando uma abordagem semelhante à da lista dourada (mas a posteriori), classificámos 431 palavras em 131 grupos (a partir de uma lista inicial de 1000 palavras distribuídas por 223 grupos), e medimos o resultado dos vários sistemas, segundo o método de Paice, como apresentado na Tabela 1-9. Os resultados deste teste mostram que os analizadores morfológicos praticamente não cometeram erros do tipo 1, contudo seu índice de erros do tipo 2 foi consideravelmente maior.

Tabela 1-9: Classificação dos sistemas segundo o método de Paice

Sistema	A	B	C	D	E	F	H
Índice de Erro Tipo 1	1	0	0	0	0	1,097E-05	3,62E-03
Índice de Erro Tipo 2	0,63	0,69	0,54	0,73	0,70	0,59	0,34

A lista usada foi o resultado de uma adaptação manual, entrando em conta com as palavras do corpus das Morfolimpíadas, da lista original, compilada pela autora do RSLP (primeira autora do presente capítulo), e que poderá, pois, inconscientemente reflectir o desempenho e o conhecimento dos problemas incluídos no próprio RSLP. Não é, no entanto, trivial obter outros agrupamentos de uma forma completamente independente da própria forma como os sistemas são desenhados, ainda que este possa ser um tópico interessante de investigação, sobretudo se for levado em conta o comportamento de utilizadores reais defrontando-se com uma base real.

5.5 Análise das semelhanças e diferenças entre os sistemas

Para além, mais uma vez, de qual o radical atribuído, várias outras análises podem ser efectuadas: Por exemplo, comparando os analisadores morfológicos entre si, que casos há em que há radicalização em alguns sistemas e não noutros? O que significam?

Em alguns casos, as diferenças devem-se a um maior número de alternativas contempladas pelo analisador morfológico em questão e podem pois ser uma medida da qualidade relativa dos sistemas; noutros casos, contudo, a diferença deve-se a diferenças teóricas no que deve ser considerado o lema (ex. o lema de *gata* é "gato" ou "gata"?, o lema de *absolutamente* é "absoluto" ou "absolutamente"? etc.). Ainda noutros, são questões de como tratar o lema de nomes próprios (ou de palavras com maiúscula inicial), como já referido.

Por outro lado, em que casos é que a radicalização é comum a todos os sistemas? Ou melhor, e mais relevante para a nossa questão inicial, em que casos é que o agrupamento é comum a todos os sistemas (independentemente da forma do radical obtida)? Apresentamos os resultados na Tabela 1-10. É interessante verificar que, embora mais semelhantes entre si, os vários analisadores morfológicos estão longe de ser equivalentes nesta tarefa. Na diagonal está o número de grupos (com dois ou mais elementos) total produzido por cada sistema.

Tabela 1-10: Quantos grupos são comuns entre os vários sistemas

	A	B	C	D	E	F	H
A	2752	1801	1335	1332	1297	1478	705
B		2930	1252	1248	1217	1729	930
C			2414	1120	1080	1204	715
D				2302	2221	985	459
E					2243	947	441
F						2566	834
H							3133

6. Conclusões

Este artigo, além de apresentar o RSLP, é pioneiro em usar os recursos criados pelas Morfolimpíadas para fazer investigação sobre outras questões, demonstrando assim a utilidade do recurso para além do possível treino de outros sistemas para avaliações posteriores.

De uma forma mais geral, estávamos preocupadas com as seguintes perguntas:

- Podemos dizer algo sobre a vantagem de usar um analisador morfológico como radicalizador?
- Podemos dizer algo sobre a vantagem de usar um radicalizador para RI em vez de um analisador morfológico?
- Podemos dizer algo sobre a avaliação de um radicalizador (casos em que perde informação) usando este tipo de comparação com um analisador morfológico?

No que respeita às perguntas 1 e 2, pensamos ter demonstrado que são programas com objectivos distintos e que, no que respeita ao agrupamento de conceitos para RI, têm potencialidades e problemas diferentes.

Pensamos também ter mostrado que a comparação em massa com analisadores morfológicos pode ser mais uma via de avaliação de radicalizadores, ainda que as componentes de frequência, não presentes no material que analisámos[8], possam ser mais um factor prático que favoreça a abordagem simples da radicalização. Como observado por Almeida & Simões (neste volume), outras serão as potencialidades de um analisador morfológico quando puder servir também análises ordenadas por frequência e/ou separar casos raros de casos mais frequentes, como fizemos na lista dourada.

Se for possível aumentar significativamente o tamanho da lista dourada (analisando mais e mais formas), e enriquecê-la com essa informação (quais as interpretações raras), poderemos refazer o algoritmo de transformação em radicalizador para que despreze as análises raras e repetir os estudos aqui apresentados.

Alternativamente, em edições posteriores das Morfolimpíadas poderíamos garantir a inclusão de um conjunto de formas nos textos para aplicar o método de Paice, podendo assim produzir uma resposta (parcial) ao problema de analisar as capacidades de derivação dos sistemas.

Dessa forma, passaríamos a poder usar as Morfolimpíadas para estudar também o desempenho de sistemas sob vários outros ângulos, nomeadamente RI e procura em corpora.

Finalmente, pensamos ter produzido alguma informação quantitativa interessante sobre a língua portuguesa, ao medir o número de formas morfologicamente ambíguas (segundo cada sistema) assim como aquelas com lemas diferentes, como apresentado na tabela 1-2.

[8] A não ser implicitamente, por não aparecerem formas raras no material.