Organização e resultados morfolímpicos

Luís Costa, Paulo Rocha e Diana Santos

1. Apresentação

Neste artigo descrevemos a primeira iniciativa de avaliação (conjunta) de avaliadores morfológicos para o português, organizada pela Linguateca.

Como referido no capítulo anterior (Santos, neste volume), uma ideia fundamental subjacente ao modelo de avaliação conjunta é que os participantes, ou seja, os responsáveis pelos sistemas a ser avaliados, colaboram, na medida das suas possibilidades, na organização da avaliação.

O artigo foca o que foi necessário fazer do ponto de vista da organização central para levar a bom porto uma organização colaborativa, e portanto distribuída, assim como as tarefas que nos coube em sorte implementar, com a anuência dos participantes: a escolha dos textos e seu empacotamento; a anonimização e uniformização dos resultados dos sistemas; e a sua classificação e consequente ordenação. Apresentamos, além disso, o recurso público resultante e tecemos algumas considerações sobre o seu uso posterior.

Muito brevemente, o objectivo das Morfolimpíadas era comparar o desempenho de vários sistemas em relação à seguinte tarefa abstracta: dada uma forma, produzir todas as variantes possíveis de análise morfológica que lhe pudessem ser atribuídas em português (correspondendo a contextos diferentes, naturalmente). Devido ao interesse de avaliação em áreas relacionadas, considerámos também duas outras tarefas: simples verificação, e radicalização ("stemming", em inglês).

Nesta comparação entre diferentes analisadores morfológicos, os pontos principais podem ser assim resumidos:

- Escolha de um conjunto de características comparáveis, a saber: categoria gramatical, lema, número, género, pessoa, tempo, grau, diminutivo/aumentativo (na Tabela 1-1 encontram-se os valores possíveis para cada categoria);
- Comparação das unidades reconhecidas pelos sistemas e sua máxima uniformização (basicamente, definir a que palavras ou unidades faz sentido associar uma classificação morfológica);
- Criação (cooperativa) de uma série de resultados correctos (a chamada "lista dourada") para serem usados como bitola de comparação;
- Obtenção de um conjunto de textos de teste de características variadas, que incluam os elementos da lista dourada;
- Comparação dos sistemas com base na lista dourada, com a produção de estatísticas várias e comparações alternativas usando diferentes pressupostos.

Noutros textos (Santos & Rocha, 2003; Santos *et al.*, 2003, Santos & Barreiro, 2004) já mostrámos que os tipos de saída e de análise dos diferentes sistemas participantes eram, de facto, diversos, como pode ser devidamente apreciado nos restantes capítulos deste volume. Esta diversidade, convém salientar, exigiu um elevado pós-processamento da nossa parte, a ponto de transformarmos cada sistema noutro de funcionamento quase irreconhecível para os seus autores. Alguns exemplos dessa remodelação serão apresentados na secção 5.

Também o processo seguido na construção da lista dourada, assim como várias das suas propriedades e limitações têm sido abundantemente documentadas: além dos artigos acima mencionados, veja-se Barreiro & Afonso (neste volume) e Baptista (neste volume). Parece-nos contudo útil referir neste capítulo de apresentação as grandes linhas metodológicas seguidas, além da apresentação do próprio resultado, na secção 3.

Além disso, discutiremos os seguintes assuntos: a selecção e criação dos textos; a escolha das métricas de avaliação; e a apresentação do recurso final, fornecendo uma caracterização preliminar.

2. Historial

Na tabela abaixo apresentamos, do ponto de vista da organização e dos participantes, a evolução temporal desta avaliação conjunta.

Tabela 1-1: Calendário das Morfolimpíadas

Fevereiro de 2002	Inscrições abertas para declaração de interesse
Junho de 2002	Apresentação da primeira proposta
Julho de 2002	Abertas inscrições para participação no ensaio
Setembro-Outubro 2002	Ensaio e encontro presencial
Março de 2003	Inscrições abertas para as Morfolimpíadas
Abril de 2003	Criação da lista dourada manual
Maio de 2003	Análise dos textos pelos sistemas participantes. Publicação da lista dourada manual
Junho de 2003	Publicação da lista dourada total. Encontro e apresentação dos resultados
Novembro de 2003	Disponibilização pública do recurso

Principiámos por efectuar um ensaio, muito útil para todas as partes. A diferença entre o ensaio e a própria avaliação conjunta é que o primeiro pretende apenas testar se os participantes perceberam o mecanismo, se os programas concebidos pela organização funcionam, e se as métricas de avaliação são justas, enquanto na própria avaliação os sistemas são avaliados e ordenados de acordo com o seu desempenho.

Entre o ensaio, documentado em Santos *et al.* (2003) e a própria avaliação houve algumas diferenças importantes, além de obviamente terem sido usados diferentes textos e diferentes listas douradas:

- A questão das unidades polilexicais, ou multipalavra, foi descartada nesta primeira edição, por ter sido impossível, em tempo útil, definir um método consensual de avaliar esta capacidade (por todos, aliás, teoricamente considerada relevante para o português)
- A questão das palavras gramaticais foi também descartada pela organização
- Algumas melhorias na forma de distribuição da colecção dos textos foram incluídas

Note-se que toda a informação relativa às Morfolimpíadas se encontra disponível em http://www.linguateca.pt/Morfolimpiadas/.

3. Lista dourada

Para comparar os sistemas, desenvolvemos um recurso de calibragem a que chamámos "lista dourada", contendo todas as análises possíveis de um conjunto de formas segundo um conjunto de directivas. Como exemplo, veja-se a forma *Caminha* com quatro análises (uma por linha):

```
Caminha÷PROP÷Caminha÷.÷S÷.÷F÷.÷.÷.÷.
Caminha÷SUB÷cama÷.÷S÷.÷F÷.÷D÷.÷.
Caminha÷V÷caminhar÷IMP÷S÷2÷.÷.÷.÷.÷.
Caminha÷V÷caminhar÷PR_I÷S÷3÷.÷.÷.÷.÷.
```

Uma lista dourada inicial foi criada no ensaio para concretizar melhor todo o processo e convergir nas opções que viriam a ser tomadas. Outra, final, foi criada para comparar os sistemas.[1]

Na presente secção, começamos por descrever os pressupostos e directivas usados na atribuição de análises às formas da colecção dourada, descrevendo em seguida pormenorizadamente a metodologia empregue para chegar ao resultado final.

[1] Teria sido interessante comparar os sistemas também com base na lista inicial (era mais um conjunto de formas cuja solução era conhecida) mas não nos lembrámos deste pormenor ao compilar os textos da segunda vez, e portanto não garantimos a inclusão de todas essas formas. Por outro lado, poder-se-ia dizer que os sistemas, ao participar no ensaio, já teriam corrigido o seu funcionamento precisamente para essas formas, que não forneceriam, portanto, uma base justa de comparação.

3.1 Directivas de classificação para cada forma

Devido às diversas classificações dos sistemas, mesmo excluindo as classes gramaticais fechadas – agrupadas numa única classe GRAM –, foi necessário tomar algumas decisões, sem dúvida contestáveis, quanto à classificação das formas constantes na lista dourada (Barreiro & Afonso, neste volume). Sendo assim, estabeleceram-se inicialmente algumas directrizes quanto à classificação das formas na lista dourada, que tiveram igualmente consequências no tratamento da saída dos sistemas:

- Levando em conta que nem todos os sistemas tratavam a derivação, optámos por incluir duas instâncias das palavras derivadas na lista dourada: uma indicando a própria forma (*encantadora*, lema "encantador"), e outra com a indicação informal da respectiva derivação (lema "encantar", sufixo "dor").
- Lusismos, brasileirismos e estrangeirismos foram devidamente assinalados. Palavras que apenas diferem na ortografia foram assinaladas com ambos os lemas (*facto*, lema "facto/fato").
- Não foram incluídas na lista dourada formas que apenas existem em português como parte de uma expressão multi-palavra (como *fortiori*).
- Optou-se por omitir o género e número das siglas e de nomes próprios quando usados como apelidos.
- Formas (infelizmente) habituais mas agramaticais (*hades* como forma do verbo *haver*) foram integradas na lista dourada, embora classificadas como inexistentes, sendo também indicada a forma correcta a que corresponderiam.[2]
- Algumas palavras completamente inventadas foram incluídas na lista dourada, classificadas como tal (é o caso de *zvlicour*).
- A classificação "rara" também foi usada para palavras ou formas pouco habituais (*augusto* como substantivo comum masculino, designando um tipo de palhaço), e algumas formas atestadas na Rede foram consideradas válidas (*ufólogo*), mesmo quando ausentes dos dicionários.[3]
- Tratámos de forma diferente a homografia acidental (por exemplo, *casa* como substantivo ou forma do verbo *casar*) da homografia sistemática intracategorial (por exemplo, no imperfeito do indicativo o português não faz diferença entre a primeira e a terceira pessoas)[4]: assim, quando a mesma realização formal corresponde na língua a um conjunto de acepções morfológicas distintas (*ajudava* pode ser primeira ou terceira pessoa do imperfeito do verbo *ajudar*)[5], essa forma é apenas contabilizada – e indicada na lista dourada – uma vez, visto que as Morfolimpíadas, convém recordar, avaliam a classificação morfológica fora de contexto.

3.2 Metodologia da compilação das várias formas

Após um processo muito moroso, no ensaio, em que cada participante no ensaio enviou para a organização cerca de trinta formas com a análise por ele considerada certa, uniformizadas por nós e depois revistas por outro participante, e em que muitos dos problemas sem consenso

[2] Note-se, a esse respeito, que também se poderia considerar outro erro: a forma em maiúsculas Hades, nome próprio referente ao deus grego dos infernos, mas não considerámos na lista dourada em nenhum caso que formas em minúsculas poderiam ser digitações erradas de nomes próprios (mais uma hipótese, embora pacífica, que convém documentar).

[3] É sabido que não há nunca uma sobreposição total entre o conteúdo de um dicionário e o uso real das palavras em texto, devido a vários fenómenos como os *hapax legomena* (palavras que apenas são usadas uma única vez) e ao desfasamento entre o léxico atestado e o presente uso da língua. Reis (1993) apresenta um estudo quantitativo preliminar sobre o português.

[4] Veja-se Santos (1996x) para uma estimativa qauntitativa dos vários tipos de ambiguidade.

[5] Note-se que, de uma forma complementar, estes correspondem aos casos mais interessante para avaliar a desambiguação morfológica em contexto, o que está fora do âmbito das Morfolimpíadas, e que é uma consequência/subproduto da análise sintáctica. Apenas dois sistemas se prontificaram a também tentar esta tarefa, que ainda foi experimenatda no ensaio, mas que devido à falta de recursos humanos e à existência de apenas dois interessados foi adiada para futuras avaliações conjuntas.

foram isolados (e consequentemente retirados da lista), seguimos, para a avaliação conjunta propriamente dita, o seguinte processo em três fases, que se mostrou também trabalhoso:

1. congregámos 380 formas (sem análise) enviadas pelos participantes e observadores;
2. seleccionámos mais 131 formas por inspecção dos textos, já coligidos através do método descrito na secção 5 com base nas 380 formas iniciais;
3. desenvolvemos um programa para extrair automaticamente casos de divergência entre os sistemas, de forma a aumentar o poder de discriminação da lista dourada resultante. (Naturalmente, este fase decorreu após a participação dos sistemas.)

O programa a que chamámos **usurário** foi desenhado para, dado um conjunto de palavras analisadas morfologicamente por um conjunto de sistemas, e normalizadas no formato universal descrito na secção 5:

- Procurar palavras onde haja maior divergência entre os sistemas
- Procurar obter exemplos com diferentes características. Por exemplo, no caso de a forma *tremendamente* já ter sido escolhida, não selecciona as formas *socialmente* ou *internamente*, cujas características morfológicas são demasiado parecidas.

Além disso, a selecção das formas é semi-aleatória. Ou seja, apesar de o programa começar pelas palavras com as análises mais divergentes, este não é o único critério que determina a escolha de uma palavra: cada palavra candidata é ainda objecto de um sorteio.

O **usurário** recorre a uma função que mede o grau de divergência global (soma das divergências entre os sistemas um a um), comparando a divergência entre as análises de uma determinada forma pelos vários sistemas. O cálculo da divergência entre dois sistemas entra em conta com o número de análises da forma por um sistema que não são contempladas pelo outro sistema e vice-versa.[6]

3.3 Caracterização da lista dourada utilizada nas Morfolimpíadas

Todas estas formas, cujas análises foram revistas por quatro a seis pessoas diferentes, produziram a lista dourada utilizada, cuja constituição é resumida na Figura 1-1.

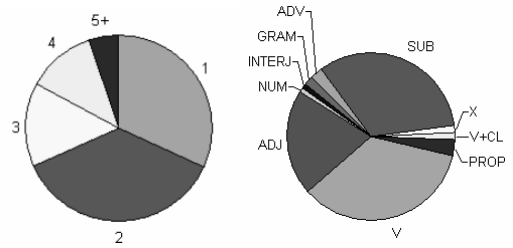

Figura 1-1: Caracterização da lista dourada (número de análises por forma e classe gramatical das análises)

Na Figura 1-2 pode ver-se uma panorâmica global da distribuição dos elementos da lista dourada nos textos usados nas Morfolimpíadas. Esta panorâmica demonstra claramente que as formas utilizadas não foram escolhidas por critérios de frequência, mas sim com a intenção de ilustrar a maior parte dos casos em que os investigadores estavam conscientes de que a análise podia ser difícil, complexa, ou não consensual.

[6] A função de comparação exclui, evidentemente, o campo `outros`, que foi por definição utilizado para colocar informação específica de cada um dos sistemas.

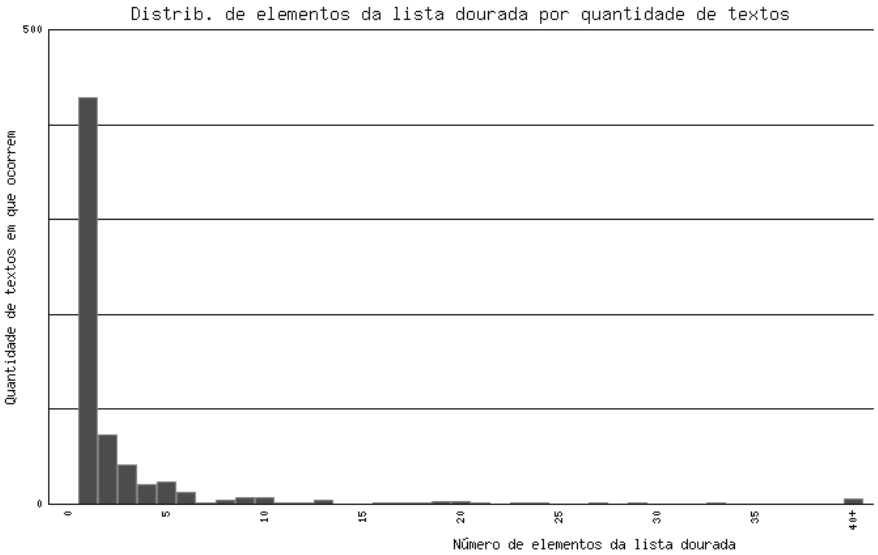

Figura 1-2: A frequência dos elementos da lista dourada nos textos

Na Figura 1-3, usamos as frequências extraídas do CETEMPúblico e do CETENFolha, e o seu análogo para as colecções de páginas Web em português WPT03 e WBR99, para ajuizar da relativa frequência das formas empregues. As colunas “Primeiros 1%” e “Primeiros 10%” contabilizam a percentagem dos elementos da lista dourada que figuram entre, respectivamente, os 1% e os 10% átomos mais frequentes ou do CETEMPúblico ou do CETENFolha (idem para a linguagem na Web). A coluna “Existente” indica as palavras que existem em pelo menos um destes corpora e a coluna “Inexistente” aquelas que não existem em nenhum deles.

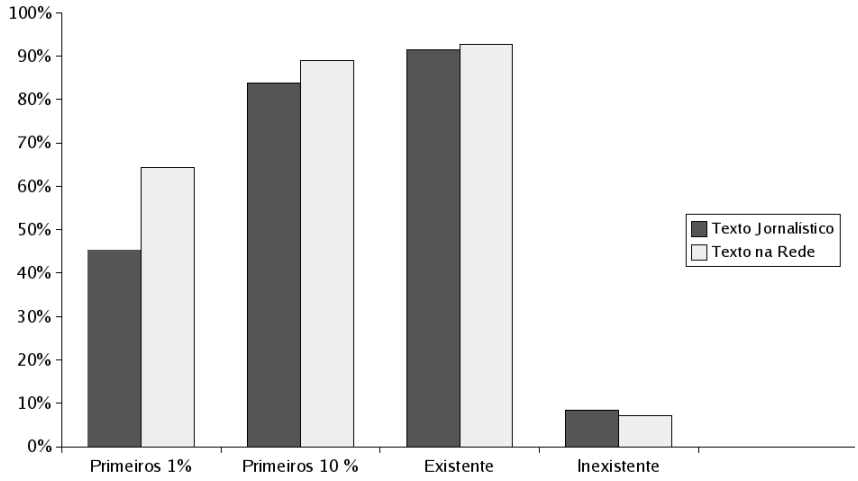

Figura 1-3: Os elementos da lista dourada em termos das estimativas de frequência em português

Para além da lista dourada, compilámos mais três conjuntos de formas interessantes para comparação, que, contudo, não foram levados em conta na avaliação dos sistemas, por falta de tempo, mas que se encontram acessíveis para posterior estudo:

- Todas as formas com hífenes: 774 casos
- Uma selecção de números e não-palavras: 247 casos
- Um conjunto de 101 palavras exclusivamente portuguesas e outras tantas exclusivamente brasileiras.

4. Criação dos textos

Aos sistemas concorrentes foi fornecido um conjunto de textos variados, em três formatos, que tinham de processar num máximo de 48 horas e devolver à organização para que a sua participação fosse válida. Nesta secção descrevemos a compilação desses textos, que teve dois objectivos:

- Esconder dos participantes quais as formas efectivamente usadas na avaliação
- Obter dados quantitativos sobre as propriedades morfológicas de textos em português

4.1 Primeira fase: vários Castores

Começámos por desenvolver um programa (a que chamámos **Castor**) que pesquisa, num corpus em formato CQP (Christ *et al.*, 1999), as palavras ou expressões que se encontram numa dada lista (no nosso caso a lista dourada), e obtém excertos contendo cada uma dessas palavras, numa posição aleatória dentro do excerto. São parâmetros do programa o tamanho dos excertos (em unidades ou frases) e restrições a partes de corpora (tal como: só texto traduzido; só texto de prosa; apenas texto não jornalístico, etc.).

O **Castor** foi aplicado a todos os corpora do projecto AC/DC (Santos & Sarmento, 2003) (incluindo o CETEMPúblico e o CETENFolha), ao COMPARA (Frankenberg-Garcia & Santos, 2003a) e à Floresta Sintá(c)tica (Afonso *et al.*., 2002b), usando 8 frases como tamanho do excerto, com um mínimo de duas frases antes e depois da palavra a esconder. Implementámos também uma variante do **Castor** para a parte da literatura infantil do corpus PAROLE (2000) e para o texto oral transcrito de Português Falado - Documentos Autênticos (2001), em que o programa selecciona, de entre os excertos com oito frases em que cada um destes corpora foi previamente dividido, um deles contendo as palavras a incluir.

Finalmente, para as palavras que não existiam de todo nos corpora, usámos textos da rede, através do motor de procura Google (com a opção de procurar apenas em páginas em português), seleccionando manualmente um contexto de tamanho aceitável. Para palavras inventadas por nós, que não encontrámos na Internet nem nos corpora, editámos textos reais de forma a introduzi-las. Recorremos também a alguns extractos de folhetos publicitários com bastantes erros, que digitalizámos.

4.2 Segunda fase: Castor-chefe

Dado esse primeiro conjunto de textos, executámos um novo programa (o **Castor-chefe**) que escolhe que extractos vão ser usados, tentando maximizar o número de origens distintas envolvidas, assim como equilibrar questões como variante e género. Na Figura 1-4 apresentamos a conjugação dos vários programas.

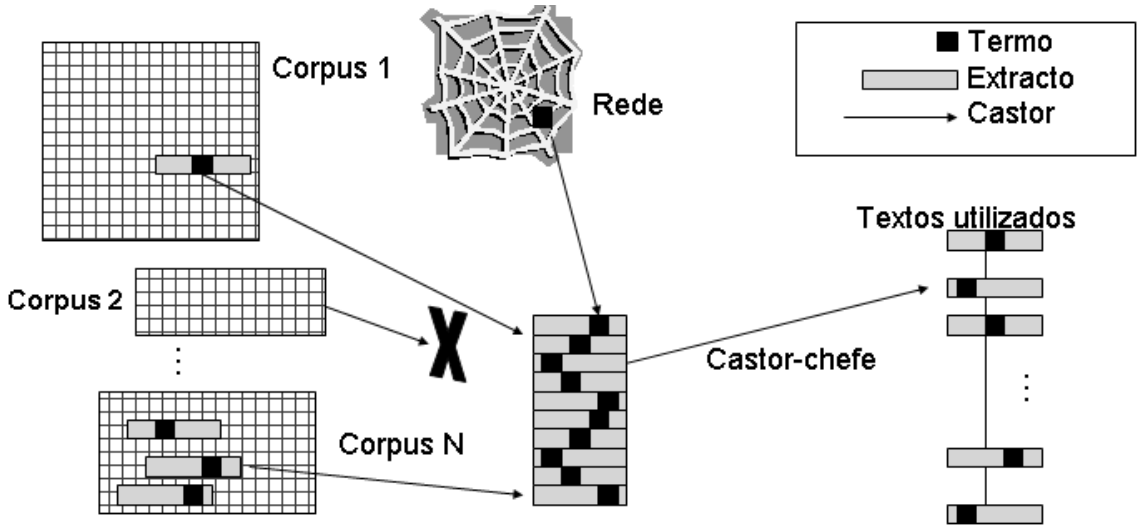

Figura 1-4: Obtenção dos textos a partir dos vários recursos textuais disponíveis

O **Castor-chefe** produziu um ficheiro com todos os textos, separados por uma marcação do tipo <TEXTO id=4>[7] e outro ficheiro mais pequeno com a informação da origem e da palavra que motivou a inclusão de cada um deles.

4.3 Preparação dos ficheiros que foram distribuídos

Esse ficheiro foi, então, sujeito a uma revisão manual, sendo alguns extractos retirados e substituídos por outros, e algumas características de formatação retiradas, dando origem ao ficheiro no formato de texto seguido: `ts.txt`.

Depois, através de alguns programas simples em Perl, na sua maioria já desenvolvidos no âmbito do projecto AC/DC, criámos os outros dois formatos, nomeadamente o de "uma unidade por linha" (`uul.txt`) e o de "um tipo só", ordenado alfabeticamente (`uts.txt`). A divisão em três formatos foi motivada pela convicção de que seria possível conseguir mais comparabilidade se a organização fornecesse uma atomização. Na Figura 1-5, apresentamos uma secção de cada um destes ficheiros.

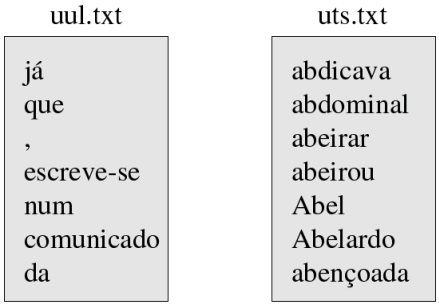

Figura 1-5 Exemplo dos ficheiros `uul.txt` e `uts.txt`

No ensaio, também editámos manualmente os ficheiros nestes dois últimos formatos, verificando casos óbvios de erros de atomização que corrigimos em ambos os ficheiros.[8] Também tentámos comparar a análise de palavras com "espaços no meio" ou expressões multipalavra, criando um ficheiro `mweuts.txt`, o análogo do `uts.txt`, mas, dado que não houve consenso quanto às métricas a usar nem quanto aos critérios a seguir (adicionado ao facto de nem todos os participantes reconhecerem esta questão como uma tarefa do âmbito estrito da morfologia), esta tarefa foi adiada para futuras edições.

O raciocínio subjacente à disponibilização dos textos em três formatos diferentes assumia que alguns sistemas tinham um módulo de atomização separado da análise morfológica propriamente dita, o que não se veio a verificar para nenhum sistema. Assim, exceptuando o caso da comparação de radicalizadores descrita em Orengo & Santos (neste volume), verificámos que esse processo induzia mais ruído e não favorecia a comparação, porque a maioria dos sistemas ainda aplicava, após a nossa, a sua própria atomização (duas atomizações seguidas), criando assim resultados artificiais e inúteis. No que se descreverá em seguida, portanto, apenas trataremos do comportamento dos sistemas aplicado ao ficheiro `ts.txt`.

4.4 Caracterização dos textos

Além da caracterização os textos conforme a variante de português usado e o meio original (que esperamos não polémica), foi também atribuída aos textos uma classificação de género: literário, jornalístico, expositivo (texto enciclopédico ou semelhante), linguagem especializada (por exemplo, textos legais) e comunicação situada (transcrições de comunicação oral). Aos textos que escapavam a qualquer destas classificações foi atribuído o género "outros".

[7] No ensaio, tínhamos agregado os textos sem identificação, apenas separados por duas mudanças de linha contíguas, o que tornou extremamente complicada a recuperação das separações de textos diferentes, necessárias para avaliar o desempenho de acordo com a variante, o género e o meio.

[8] A maior parte destes erros era devida à introdução da linguagem oral no conjunto dos textos, que usa uma pontuação diferente daquela para a qual os programas foram originalmente concebidos, mas também foram corrigidos manualmente casos combinando parênteses e algarismos.

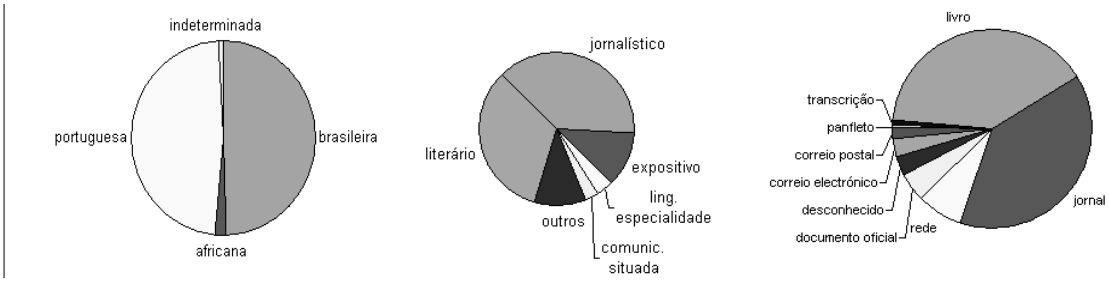

Figura 1-6: Repartição dos textos, por variante, género, e meio original

~~Além da caracterização os textos conforme a variante de português usado e o meio original (que esperamos não-polémica), foi também atribuída aos textos uma classificação de género: literário, jornalístico, expositivo (texto enciclopédico ou semelhante), linguagem especializada (por exemplo, textos legais) e comunicação situada (transcrições de comunicação oral). Aos textos que escapavam a qualquer destas classificações foi atribuído o género "outros".~~

A distribuição dos textos usados nas Morfolimpíadas com base nessa caracterização encontra-se na Figura 1-6. Na Tabela 1-2, apresentamo~~s algumas~~s outras estatísticas sobre os excertos usados.

Tabela 1-2: Panorâmica dos textos usados

Número de textos distintos	613
Origens (corpora) diferentes	18
Tamanho máximo em unidades	344
Tamanho mínimo em unidades	27
Tamanho médio em unidades	131,9

5. Processamento da saída dos sistemas

Para lidar com o facto de os sistemas participantes nas Morfolimpíadas apresentarem diferentes formatos como resultado, criámos uma série de programas que convertiam as diferentes saídas para um formato universal que permitisse a comparação mútua. Na Figura 1-7 encontra-se uma descrição gráfica do processo a que a saída de cada sistema foi submetido.

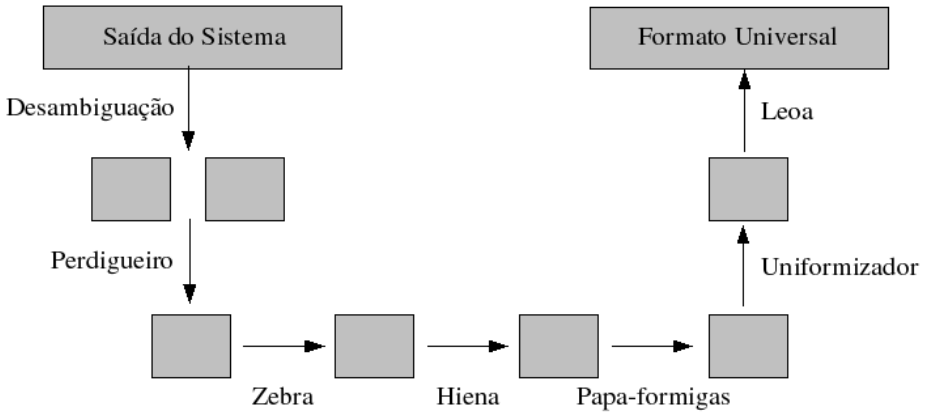

Figura 1-7: Conversão dos resultados dos analisadores morfológicos

Como referido acima, os diferentes sistemas produziram saídas bem diferentes. Abaixo apresentamos como exemplo diversas análises da palavra *aparece*:

```
aparece          [aparecer] V IMP 2S VFIN        [aparecer] V PR 3S IND VFIN

aparece: pode ser obtida do lema aparecer
```

```
aparece lex(aparecer,[CAT=v,T=p,TR=i,P=3,N=s]), lex(aparecer,
[CAT=v,T=i,TR=i,P=2,N=s])

{aparece,aparecer.V:P2's:P3s:Y2s}

'aparece'.
[ 'aparecer', 'CAT','v', 'NUM','s', 'GEN',_, 'PES','3', 'MOD','ind',
'TMP','pre'].
[ 'aparecer', 'CAT','v', 'NUM','s', 'GEN',_, 'PES','2', 'MOD','itfp'].

aparece=<V.[INT.TI.][PRES.ELE.IMPER-AFIRM.TU.]N.[a.em.][aparecer]0.>
```

Os principais programas desenvolvidos neste contexto foram (um para cada sistema):

Desambiguação: para criar duas versões dos resultados, no caso do único sistema que produzia atomizações alternativas: uma com atomização com o maior número de unidades (excluindo separação de hífenes) e outra com o menor número de unidades. Por exemplo, *em geral* foi definido no primeiro caso como um advérbio com duas palavras, analisadas separadamente no segundo caso.

Perdigueiro: para repor a identificação dos textos (a já mencionada marcação TEXTO) que alguns sistemas trataram como fazendo parte dos próprios textos.

Zebra: para traduzir o formato próprio do sistema para o formato universal, ou seja, um ficheiro com as análises de cada forma separadas pelo caracter ×. O conjunto de análises de cada forma é por seu turno apresentada assim (as denominações dos valores são iguais para todos os sistemas, e não coincidem com a nomenclatura de nenhum deles):

```
×
forma÷PoS÷lema÷tempo÷número÷pessoa÷género÷grau÷D/A÷outros
forma÷PoS÷lema÷tempo÷número÷pessoa÷género÷grau÷D/A÷outros
×
```

A Tabela 1-3 lista os valores usados. Além disso, é usado um ponto (“.”) para indicar que o campo em questão não se aplica à análise efectuada (por exemplo, o tempo para um advérbio). No caso de contracções e verbos com clíticos, é usado o sinal “+” para unir os valores respeitantes aos diversos componentes (por exemplo, a forma *escrevi-a* terá a categoria gramatical (abreviadamente, PoS, do inglês “Part of Speech”) “V+CL”, pessoa “1+3” e número “S+S”).

Tabela 1-3: A informação em questão nas Morfolimpíadas

PoS	ADJ, ADV, GRAM, INTERJ, NUM, PROP, SUB, V, V+CL (verbo+clítico)
Tempo	COND, FT_C, FT_I, GER, INF, INFP, IMP, PP, PR_C, PR_I, PSI_C, PSI_I, PSP_I, PSM_I
Número	S, P
Pessoa	1, 2, 3
Género	F, M, I (invariável)
Grau	COMP, SUP
D/A	AUM, DIM
Outros	ABREV, CARD, CONTR, FRAC, LETRA, ORD, QUIM, SIGLA

Por exemplo, a análise da palavra *comando* de um dos sistemas foi transformada no seguinte:

`comando [comando] N M S [comandar] V PR 1S IND VFIN`

`×`
`comando÷SUB÷comando÷.÷S÷.÷M÷.÷.÷.`
`comando÷V÷comandar÷PR_I÷S÷1÷.÷.÷.÷.`
`×`

Hiena: para modificar o processamento de algumas questões de diferente atomização de forma a tornar a comparação mais justa. Foram desenvolvidas hienas:

- para os clíticos (de forma a juntar a informação V+CL dos sistemas que consideram verbo e clíticos como átomos diferentes),
- para as contracções (no caso dos sistemas que as separam) e
- para os números (no caso dos sistemas que apresentam cada dígito separado).

Também criámos algumas hienas menores que tinham a ver com problemas de atomização muito específicos, como por exemplo abreviaturas e pontuação.

Abaixo, apresentamos o caso de um sistema que analisou *abaixando-se* em dois componentes, e o resultado da execução da respectiva hiena:

```
×
abaixando-÷V÷abaixar÷GER÷.÷.÷.÷.÷.÷<hyfen>
×
se÷PR_PES÷se÷.÷I÷3÷I÷.÷ACU÷<refl>
×
```

```
×
abaixando-se÷V+CL÷abaixar+se÷GER÷.+I÷.+3÷.+I÷.÷.÷.÷.
×
```

Note-se que, ainda assim, algumas disparidades entre os diferentes sistemas se mantiveram, provocadas pelas análises distintas dessas palavras gramaticais. Por exemplo, *reuniram-se* teve como lemas “reunir”, “reunir-se” e “reunir-me”, conforme os sistemas analisaram separadamente os dois componentes ou não, e conforme o lema atribuído ao clítico.

Papa-formigas: para transformar as várias análises gramaticais que um analisador tivesse atribuído a uma dada forma em exactamente **uma** análise com a categoria GRAM, visto que foi decidido não comparar as análises gramaticais.

Uniformizador: para lidar com o caso das formas sistematicamente ambíguas que todos os sistemas reconhecem "em bloco", transformando o conjunto de análises em apenas uma, de forma a não dar mais peso a este tipo de formas. Um exemplo são os infinitivos pessoais das primeira e terceira pessoas do singular e o infinitivo impessoal, que correspondem sempre à mesma forma em português.[9] No exemplo abaixo, apenas a primeira análise foi mantida, para concordar com a lista dourada, também construída obedecendo a este critério.

```
×
reunir÷V÷reunir÷INF÷.÷.÷.÷.÷.÷.
reunir÷V÷reunir÷INFP÷S÷1÷.÷.÷.÷.
reunir÷V÷reunir÷INFP÷S÷3÷.÷.÷.÷.
×
```

Leoa: para identificar as formas por texto, variante, meio e género (ver Figura 1-6), de forma a mais tarde se poder separar os resultados em função desses parâmetros. Este é o formato final usado para calcular os resultados, e publicamente disponibilizado aos interessados.

```
×
comando÷SUB÷comando÷.÷S÷.÷M÷.÷.÷.÷tx=1÷va=bras÷ge=ind÷me=web
comando÷V÷comandar÷PR_I÷S÷1÷.÷.÷.÷.÷tx=1÷va=bras÷ge=ind÷me=web
×
```

Além disso, desenvolvemos ainda mais dois programas, usados para comparar, respectivamente, os sistemas como radicalizadores e como verificadores:

Lagartixa: uma versão abreviada da **leoa**, que apenas apresenta cada forma analisada e respectivo lema.

Condor: que indica, para cada forma analisada, se o sistema a reconhece ou não (veja-se secção 8.1).

[9] Tivemos evidentemente o cuidado de não incluir verbos por alguns considerados defectivos em português, precisamente para evitar a questão de comparar casos em que alguns sistemas só considerariam S 3 e não também S 1.

6. Resultados

Os resultados que deram origem à definição de um vencedor e à ordenação dos sistemas foram determinados pela comparação das saídas dos diferentes sistemas com as análises presentes na Lista Dourada.

As medidas utilizadas foram a precisão e a cobertura, para conjuntos diferentes de traços: usámos todos os campos menos o campo `outros`, todos menos os campos `lema` e `outros`, só `lema` e `pos`, só `pos`, etc. "Precisão" foi definida como a percentagem de análises[10] concordantes com a lista dourada produzidas pelo sistema que o sistema produziu, "cobertura" como a percentagem de análises constantes na lista dourada, que o sistema produziu.

Apresentamos os resultados para o subconjunto da lista dourada escolhido por pessoas (Tabela 1-4) e para a lista dourada completa (Tabela 1-5), enriquecida com as formas escolhidas automaticamente pelo **usurário** (mas classificadas pela organização). Visto que este programa escolhia casos de divergência, onde por definição algum sistema tinha portanto uma resposta errada, o desempenho global naturalmente diminuiu em relação à lista dourada inicial.

Nestas e em tabelas subsequentes, apresentamos os sistemas descritos por letras, arbitrariamente atribuídas (que variam entre as tabelas). Isto deve-se a termos garantido, de forma a atrair o maior número possível de participantes, que os resultados das Morfolimpíadas seriam publicados sem indicação do sistema a que correspondiam, excepto no caso do sistema vencedor.

Tabela 1-4: Desempenho dos sistemas, desprezando os campos `lema` e `outros`, para a lista dourada manual

Sistema	A		B		C		D		E		F	
	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.
Formas comparadas	487	511	465	511	458	511	506	511	499	511	499	511
Análises lista dourada	938	991	889	991	880	991	981	991	968	991	973	991
Análises do sistema	835		855		840		780		735		821	
Análises comuns	721		711		697		643		587		630	
Precisão	86,35%		83,16%		82,98%		82,44%		79,86%		76,74%	
Cobertura relativa	76,87%		79,98%		79,75%		66,02%		61,06%		65,15%	
Cobertura absoluta	72,75%		71,75%		70,33%		65,68%		59,23%		66,89%	

A primeira coisa a salientar é a questão das dificuldades de atomização (ver secção 7.1), que levaram a que nenhum sistema reconhecesse – e portanto classificasse – a totalidade das formas incluídas na lista dourada. Donde, optámos por desdobrar a cobertura em duas: **relativa**, em que entram para as contas apenas as formas da Lista Dourada que são reconhecidas pelo sistema, e **absoluta**, em que todas as análises presentes na lista dourada são levadas em consideração.

Tabela 1-5: Desempenho dos sistemas, desprezando os campos `lema` e `outros`, para a lista dourada total

Sistema	A		B		C		D		E		F	
	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.
Formas comparadas	628	655	599	655	590	655	648	655	639	655	640	655
Análises l. dourada	1199	1255	1126	1255	1110	1255	1247	1255	1219	1255	1237	1255
Análises sistema	1056		1080		1065		979		879		968	
Análises comuns	903		886		871		808		676		754	
Precisão	85,51%		82,04%		81,94%		82,53%		76,91%		77,89%	
Cobertura relativa	75,31%		78,60%		78,38%		64,80%		55,55%		60,95%	
Cobertura absoluta	71,95%		70,60%		69,40%		64,38%		53,86%		60,08%	

[10] De notar que, ao considerarmos apenas os campos `forma` e `lema`, ou `forma` e `pos`, estamos a reduzir o número de análises distintas consideradas para comparação (por exemplo, *fui* como verbo IR e *fui* como verbo SER, sem lema, transformam-se numa análise única de categoria gramatical V).

Na Figura 1-8 apresentam-se os resultados correspondentes, para os seis sistemas, com a lista dourada total.

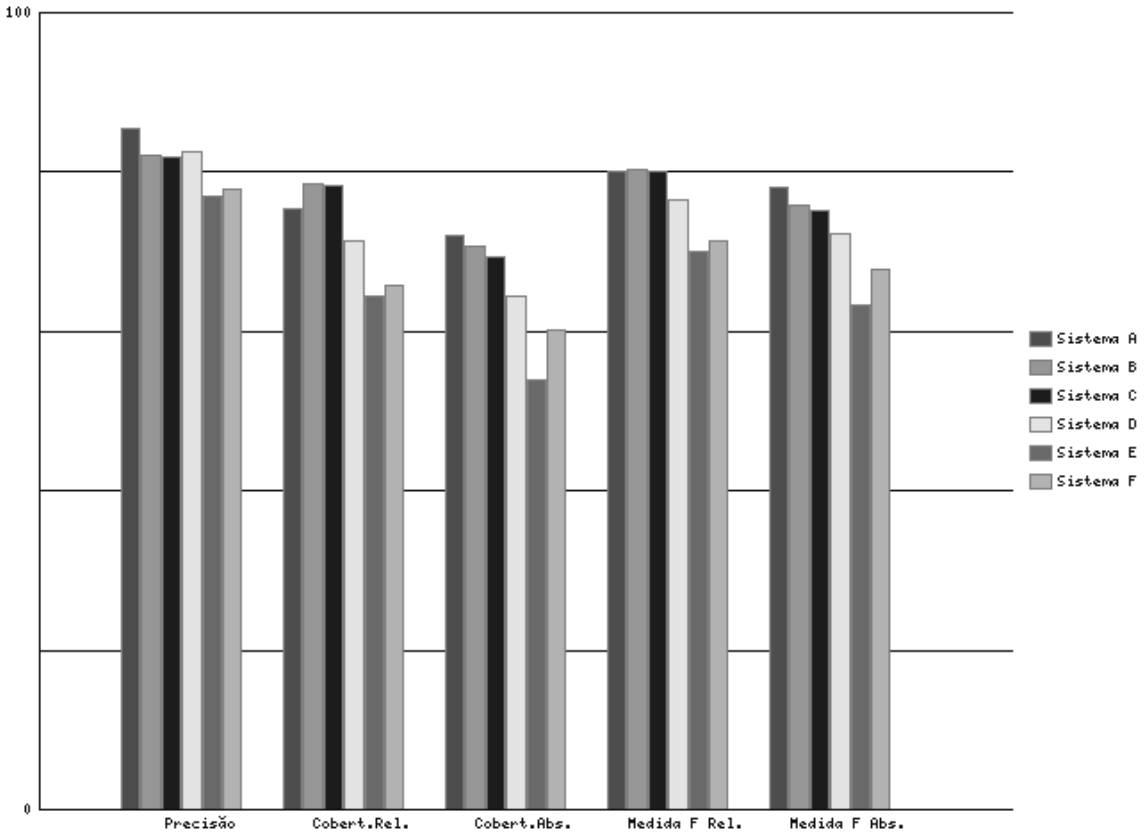

Figura 1-8: Precisão, cobertura e medida F relativas e absolutas dos seis analisadores morfológicos em comparação com a lista dourada total

O sistema A, o PALMORF (Bick, neste volume), obteve os melhores resultados em todas as medidas – exceptuando a cobertura relativa, em que o sistema B, seguido de muito perto do sistema C, obteve o melhor resultado. É, no entanto, de salientar a relativa comparabilidade de todos os sistemas, que diferiram em precisão – seguindo o processo de avaliação das Morfolimpíadas – em menos de 10%.

6.1 Critérios de comparação

As diferenças entre os sistemas, contudo, levaram-nos a testar e comparar várias funções de comparação, para averiguar se não estaríamos a prejudicar ou favorecer indevidamente algum sistema.

Assim, as seguintes experiências foram executadas antes de publicar os resultados finais, através da implementação de diversas opções nos programas de comparação, também públicos:

- ignorar as análises de interjeições
- ignorar as análises de derivação
- ignorar a análise de nomes próprios
- ignorar a comparação com as análises consideradas “raras” (e como tal assim marcadas) da lista dourada

Nenhuma destas opções alterou a ordem relativa dos sistemas. Em muitos casos – por exemplo a questão da derivação – nem sequer revalorizava positivamente o desempenho dos sistemas que tentava “proteger”. Isso evitou-nos problemas na decisão de quais os critérios a utilizar, deixando em aberto para o público em geral a definição no futuro de outras formas de comparação. Sendo assim, não foi feito (ainda) um estudo estatístico para identificar quais os factores mais relevantes, que critérios alternativos podiam ser utilizados, ou que número de

formas deveria ser retirado ou adicionado à lista dourada para mudar a ordem entre os sistemas. Qualquer pessoa interessada em morfologia, avaliação e relevância estatística o poderá fazer, contudo, dado o recurso ser disponibilizado publicamente.

7. Recurso

A nosso ver, um dos principais resultados das Morfolimpíadas foi a construção e disponibilização de um recurso único para estudar a morfologia do português e ajuizar o que falta fazer quanto à aplicação de analisadores morfológicos a texto real em português.

Além de disponibilizar os textos usados, as saídas dos programas participantes, e os próprios resultados, a Linguateca distribuiu também os programas que foram utilizados no cálculo dos resultados das Morfolimpíadas, diferindo apenas dos originais no estritamente necessário para permitir a anonimização dos sistemas participantes. Assim, contribuímos para:

- a replicação dos resultados por investigadores independentes, no sentido de permitir uma avaliação independente;
- a experimentação com as medidas e a reengenharia da classificação;
- a obtenção de melhores programas e um maior envolvimento da comunidade.

Nesta secção apresentamos os primeiros estudos feitos com base nesse recurso, apresentando ao mesmo tempo uma melhor caracterização do material disponibilizado.

7.1 Atomização

Uma das conclusões mais avassaladoras deste trabalho foi a enorme disparidade entre as técnicas e heurísticas de atomização, ou seja, de identificação de unidades sobre as quais um analisador morfológico iria depois trabalhar. Embora já previssemos diferenças significativas (Santos, 1999a), os resultados foram ainda mais pronunciados, talvez por termos usado um leque muito variado de tipos de texto.

A partir das saídas dos sistemas normalizadas no formato universal, foi possível determinar algumas medidas de concordância. Começou-se por determinar que o número total de unidades diferentes em `ts.txt` identificadas pelos sistemas foi de 21411. Quanto ao número de unidades reconhecidas simultaneamente por todos os sistemas, ficou-se pelos 13545.

Em seguida, procurámos verificar se algum dos sistemas era particularmente diferente dos outros, já que isso poderia comprometer a obtenção de uma visão de conjunto. A Tabela 1-6 mostra os números máximo e mínimo de unidades identificadas em comum considerando todos os sistemas dois a dois, três a três, etc. Nela podemos comprovar que os sistemas são consideravelmente diferentes entre si no que respeita à atomização, mas não existe um que se destaque nesse aspecto em particular.

Tabela 1-6: Unidades diferentes (tipos) identificadas em comum

Sistemas	Máximo	Mínimo
2 a 2	16 713	13 827
3 a 3	15 279	13 640
4 a 4	14 938	13 568
5 a 5	13 705	13 545
os 6	13 545	

No total, o número de formas únicas (ou seja, só reconhecidas por um dos sistemas) ascendeu a 2021. As formas seguintes – separadas por vírgula seguida de espaço – ilustram algumas das diferenças de atomização observadas. Estes exemplos pretendem apenas demonstrar casos raros e não estabelecer uma panorâmica das diferenças de atomização:

```
em homenagem ao, África do Sul, DOS Ovos De Ouro João Impaciente, sistema
nervoso, nipo, inovação, EXPO 98, Toyota Ipsum, Príncipes Yuriko, de estudo,
logo-, 79, música pop, Tuna, em nome dele, 22.30h , uma porcaria, Linhas de
Torres, al., Apesar destas
```

Estudámos também a concordância entre os pares `forma-pos` e `forma-lema` e trios `forma-pos-lema` identificados por cada um dos sistemas. A Tabela 1-7 inclui também a percentagem dessas unidades que foi reconhecida por todos, bem como, para cada caso o total

de unidades distintas considerando todos os sistemas (“Totais”) e o número de unidades consensuais (“Comuns”) que é utilizado para calcular a percentagem de concordância.

Tabela 1-7: Concordância para alguns pares de análises

Sistema	A	B	C	D	E	F	Comuns	Totais
Unidades dif.	16 850	17 448	17 304	17 044	15 694	16 357	13545	21411
Concordância (%)	80,39	77,63	78,28	79,47	86,31	82,81		
`Forma x PoS` dif.	19 861	21 456	20 639	20 395	20 144	20 729	10881	34272
Concordância (%)	54,79	50,71	52,72	53,35	54,02	52,49		
`Forma x Lema` dif.	19 116	20 498	19 229	19 818	18 595	19 244	9661	35457
Concordância (%)	50,54	47,13	50,24	48,75	51,95	50,20		
`Forma x PoS x Lema` dif.	20 537	21 725	20 944	20 514	20 489	21 077	9239	40580
Concordância (%)	44,99	42,53	44,11	45,04	45,09	43,83		

Muitos outros dados poderiam ser extraídos da comparação das saídas dos sistemas e, de facto, essa é uma das razões por que consideramos tão importante o termos disponibilizado essas mesmas saídas, devidamente anonimizadas, publicamente.

Enquanto a Tabela 1-8 produz um resumo quantitativo da forma gráfica das unidades produzidas por cada sistema, a Tabela 1-9 indica quantas formas têm (pelo menos) uma análise cuja classificação gramatical é SUB, ou V, etc.

Tabela 1-8: Características das unidades identificadas por cada sistema

Sistema	A	B	C	D	E	F
Só letras	68430	66923	63501	67857	66517	68517
Letras com hífenes	69150	67608	64364	68651	67237	68579
Com letras e dígitos	0	47	65	45	0	48
Com letras e pontuação	7	154	249	255	7	1168
Só dígitos	912	624	508	581	912	566
Dígitos com pontuação	0	89	141	112	0	96
Só pontuação	13387	12745	12235	12599	13388	11826
Unidades simples	83386	80275	74703	82289	80044	77825
Unidades com várias palavras	2	1090	2900	1	1432	1343

Tabela 1-9: Categoria gramatical das formas

Formas com análise	A	B	C	D	E	F
Siglas	203	32	0	730	200	97
SUB	32211	17053	28491	27269	30970	17621
V	17939	17117	16186	16968	17344	15982
V+CL	609	607	606	598	609	562
ADJ	9171	7845	7693	11084	8535	6642
ADV	4505	3474	7495	6937	4720	3874
NUM	3089	1205	2289	2828	288	2486
PROP	1225	1441	2880	3627	195	4206
INTERJ	996	0	932	0	103	153
GRAM	29390	28802	27546	30625	4167	28783
PONT	13387	12482	12150	10897	2196	11804
Contracção	5811	5548	5300	5713	1447	5463

A Tabela 1-10 ilustra a ambiguidade morfológica reconhecida pelos diversos sistemas – em termos de número de análises por forma. Pode, portanto, afirmar-se que mais de 25% das palavras que ocorrem em textos em português são ambíguas morfologicamente.

Tabela 1-10: Formas por número de análises

Formas com exactamente	A	B	C	D	E	F	min	max	média
uma análise	63,1%	83,6%	62,7%	62,6%	59,6%	80,5%	59,6%	83,6%	68,7%
duas análises	28,4%	12,3%	30,5%	27,1%	29,0%	15,3%	12,3%	30,5%	23,8%
três análises	5,8%	3,2%	5,4%	9,5%	7,8%	3,3%	3,2%	9,5%	5,8%
quatro análises	2,1%	0,7%	1,1%	0,5%	2,9%	0,8%	0,5%	2,9%	1,4%
mais de quatro análises	0,7%	0,2%	0,4%	0,2%	0,7%	0,1%	0,1%	0,7%	0,4%

A Tabela 1-11 detalha que tipos de ambiguidade são detectados pelos sistemas: palavras que têm uma acepção gramatical e outra não gramatical são as mais frequentes, seguidas, numa faixa de frequências similar, por ambiguidade intra-verbal e verbo-substantivo.[11]

Tabela 1-11: Formas com análise ambígua

Formas com ambiguidade	A	B	C	D	E	F	min	max	média
SUB/ADJ	16,0 %	10,3 %	11,1 %	4,9 %	10,2 %	18,5 %	4,9%	18,5%	11,8 %
SUB/ADV	5,3 %	0,7 %	2,8 %	2,5 %	3,9 %	1,2 %	0,7%	5,3%	2,7 %
SUB/V	16,5 %	27,0 %	17,1 %	11,8 %	14,2 %	21,9 %	11,8%	27%	18,1 %
SUB/SUB	2,6 %	1,8 %	1,5 %	7,0 %	4,7 %	5,3 %	1,5%	7%	3,8 %
V/V	19,3 %	37,5 %	16,9 %	15,5 %	20,9 %	21,8 %	15,5%	37,5%	22,0 %
V/ADJ	5,8 %	8,4 %	6,1 %	7,3 %	6,4 %	5,8 %	5,8%	8,4%	6,7 %
ADJ/ADJ	0,1 %	0,5 %	0,0 %	0,0 %	0,3 %	2,0 %	0%	2,0%	0,5 %
ADV/ADV	0,0 %	0,0 %	0,2 %	0,0 %	0,1 %	0,0 %	0%	0,2%	0,0 %
ADJ/ADV	0,8 %	0,6 %	1,8 %	2,2 %	0,9 %	1,0 %	0,6%	2,2%	1,2 %
GRAM/outro	33,6 %	13,3 %	42,5 %	48,6 %	38,5 %	22,6 %	13,3%	48,6%	33,2 %

Finalmente, a Tabela 1-12, considerando as análises separadamente, descreve a sua interpretação a nível da categoria gramatical. (Recorde-se que uma forma pode ter uma ou mais análises, como mostrado na Tabela 1-10).

Tabela 1-12: Categoria gramatical das análises

Análises como	A	B	C	D	E	F	min. %	max. %
V	24565	22907	21295	22211	23768	19574	18,3%	22,0%
V+CL	2773	1498	1313	833	2773	1226	0,1%	2,2%
SUB	33228	20331	28982	29828	31919	18631	16,7%	27,0%
ADJ	9216	10159	9984	11094	8580	8743	6,7%	9,1%
ADV	4506	3474	7554	6937	4737	3874	3,3%	6,5%
GRAM	29390	28802	27546	30625	28347	28783	23,5%	28,5%
INTERJ	996	0	932	0	915	153	0,0%	0,8%
PROP	1378	1481	3617	3627	1359	4208	1,1%	4,2%
NUM	3089	1205	2302	2828	3027	2697	1,2%	2,5%
Total	109141	84645	101242	107983	104934	85951		

8. Três facetas das Morfolimpíadas

De forma a congregar o maior número de participantes que tivessem sistemas relacionados com a análise morfológica, anunciámos, na chamada à participação, que verificadores ortográficos e lematizadores (ou radicalizadores) também seriam bem-vindos.

Visto que tivemos a participação do br.ispell na primeira categoria (veja-se Karpischek, neste volume) e do RSLP (Orengo & Santos, neste volume) na segunda, para além de cinco

[11] Exemplos do primeiro caso (GRAM/outro) são: *a/o/e* como nomes (de letras), e outras palavras gramaticais frequentes como *como*, *pelo*, etc. que são homónimas com formas de verbo ou nome. Verifica-se a pertinência de não pesar as formas usadas pelas Morfolimpíadas pela sua frequência, ou seja, estes casos, embora muito frequentes, não são particularmente relevantes para testar um analisador morfológico

sistemas que fazem análise morfológica propriamente dita (PALMORF, jspell, ReGra, LabEL-INTEX e Smorph), podemos considerar que as Morfolimpíadas tiveram três facetas.

O presente capítulo foi dedicado à tarefa mais complicada, que aliás congregou o maior número de participantes, nomeadamente, a que se referiu à análise morfológica propriamente dita, e remetemos os leitores para os artigos referidos no caso das outras tarefas.

8.1 Os sistemas como verificadores

Mostramos, contudo, na Error: Reference source not found as capacidades de reconhecimento dos sistemas participantes, considerando-os como verificadores ortográficos (no caso dos sistemas que indicam, de alguma forma, que uma dada forma não é reconhecida – quer por ser rejeitada, quer por ser adivinhada pelo sistema).[12] Mais uma vez se confirma aquilo que apontámos acima, na secção 7.1, ou seja, que um mesmo texto é tratado diferentemente por cada sistema, logo na primeira tarefa de identificar as unidades que contém

Tabela 1-13: Resumo dos sistemas como verificadores

Sistema	A	B	C	D	E	F
Unidades identificadas	83388	81365	79272	82290	81476	79168
Unidades desconhecidas	2374	1677	1166	1057	2366	1121
Unidades adivinhadas	0	3969	2373	0	0	2302

9. Comentários finais

Não obstante as muitas simplificações feitas ao longo deste processo, assim como a tomada de algumas decisões essencialmente arbitrárias no que respeita à classificação de alguns itens menos consensuais da lista dourada, pensamos ter conseguido coligir um recurso muito interessante para o estudo da morfologia computacional do português. Foi, além disso, extremamente compensador ver tantos investigadores a partilhar connosco o seu engenho e a sua competência para chegar a um resultado comum.

Muitas diferenças entre o desempenho dos sistemas têm a ver com a divisão em módulos conceptuais que cada sistema usa para atacar o problema da análise de linguagem natural, e só podem, eventualmente, ser avaliadas e discutidas no âmbito da compreensão total de textos em linguagem natural, como exemplificado já em Santos & Gasperin (2002) e Santos (2003). Ou seja, existe um limite superior no consenso possível, que não é sensato desprezar.

Os erros, as incorrecções, e as diferenças entre variantes são factores que devem ser levados em conta, mas não é óbvio se o seu processamento (e como tal a avaliação do resultado) deva ser feito antes ou depois de uma análise morfológica.

Quanto a questões que podiam ter sido diferentemente tratadas, além de todas as sugestões que serão feitas ao longo dos outros capítulos referentes às Morfolimpíadas, parece-nos claro que um complemento, ou pelo menos uma verificação, deveria ser feita ao nível do estudo empírico da língua, depois de coligir heuristicamente um conjunto de formas para a lista dourada.

Ou seja, deverá ser possível efectuar uma avaliação independente do peso de cada forma constante na lista dourada, quer em termos de representatividade do "problema" que ilustra, quer em termos daquilo que representa no panorama descrito pela lista dourada como um todo.

Esperamos assim que o recurso disponibilizado, possivelmente associado a uma extensão ou reformulação do **usurário** ou doutro programa semelhante, permita produzir uma lista dourada mais extensa, sistemática, e de qualidade comprovada e comprovável, quer para

[12] Um leitor incauto, inconsciente das várias opções relativas ao reconhecimento e análise morfológica, poderá imaginar que verificação é um passo inicial sem o qual não poderá em seguida haver análise. Isso, contudo, é apenas uma forma de pensar e implementar. De facto, praticamente todos os analisadores morfológicos (mas não os verificadores) tentam analisar palavras "desconhecidas" ou "estranhas". A diferença está entre se o resultado do sistema mesmo assim indica que é uma análise criativa, ou não. No primeiro caso, podemos medi-lo como verificador (só para construir a tabela), no segundo não, visto que o sistema produz sempre uma análise, e não se pronuncia sobre se faz ou não parte da língua portuguesa.

futuras edições das Morfolimpíadas, quer simplesmente para teste e melhoria de qualquer analisador morfológico, presente ou futuro.