Parte III

José Mariano Gago: para documentar o "Ministro da Língua"

Diana Santos, linguateca e professora catedrática, Universidade de Oslo

Diana Santos

QUE melhor tributo poderia ser dado pelo projeto que nasceu da iniciativa e do amor de Mariano Gago pela língua portuguesa, do que aplicar as suas ferramentas e o seu saber-fazer para criar um recurso inovador que tenha o seu nome, e sobretudo, que leve a que outras pessoas venham a conhecer, no futuro, se bem que parcialmente, este génio da política científica?

Este texto apresenta um recurso sobre Mariano Gago e, ao fazê-lo, dá também a conhecer a Linguateca a um público mais vasto, que não sabe ou já não se lembra de que em 1999 em Portugal houve várias consultas aos cientistas para definirem em conjunto o futuro da sua área, organizadas por Mariano Gago. E uma das áreas em plenário, devido ao empenho dele pela valorização do português, foi precisamente "o processamento computacional da língua portuguesa". Que, por sua vez levou à criação da Linguateca, uma rede de recursos e informação sobre a nossa língua, que criaria e organizaria avaliações conjuntas para fomentar o progresso acelerado da área. Essa rede tinha vários pólos em vários países. Contudo, após mudanças governamentais e percalços administrativos que não interessa aqui mencionar, desde 2011 não tinha (ou tem) qualquer financiamento, excluindo o importantíssimo apoio informático da FCCN no alojamento dos servidores.

A Linguateca continuou, apesar disso, como uma rede de investigadores empenhados em manter e continuar a atuação em prol da língua, com uma flexibilidade que deriva de a Linguateca ter sido concebida como uma infraestrutura, e não uma "instituição", de forma a sermos livres da burocracia e das mudanças nos órgãos do poder.

Essa foi uma opção minha, aliás criticada e posta em questão pelo Mariano Gago, que confiava nas instituições, e criou várias, aliás, para continuarem a fazer o trabalho que ele desejaria. Contudo, penso que é lícito afirmar que ele ainda confiava mais nas pessoas, e que era da opinião de que nada é melhor do que deixá-las, embora aconselhadas, fazerem o que quiserem. E por isso "deixou", ou pelo menos não insistiu em criar mais uma "instituição", também porque não desdenhava das capacidades de organização pragmáticas e dinâmicas que surgem em momentos dados, e que podem complementar as iniciativas legisladas, planeadas e dirigidas de cima.

Por essa razão, ao sermos confrontados com a notícia da sua morte, fazer um gesto que não estava no programa demonstrou as vantagens dessa postura da Linguateca, de não ter de prestar contas burocráticas a ninguém, apesar da falta de financiamento e de enquadramento científico-político.

É preciso, contudo, confessar que só agora estou a ver a questão sob esse ângulo. A decisão de construir o corpo Mariano Gago como homenagem, para reverenciar a memória dele tornando esse recurso útil para outras atividades, embora sem qualquer financiamento ou plano, impôs-se como a única resposta positiva a uma calamidade tão profunda. Penso não exagerar em afirmar que a notícia da sua perda e as consequências que traria para a investigação e para o desenvolvimento do país (e, no nosso caso particular, para o futuro da língua portuguesa, que não se esgota de forma alguma em Portugal) colocou todos quantos o conheciam em choque coletivo, e a construção deste corpo revelou-se como uma espécie de catarse, um desafio ao esquecimento.

O presente texto está assim estruturado: uma breve descrição do corpo Mariano Gago, e da sua existência na ecologia da Linguateca, para que leitores que não conheçam o nosso trabalho ou a própria existência da Linguateca possam ficar com uma ideia; seguido de um capítulo em que se explica como usá-lo, apelando em seguida a uma colaboração para um seu desenvolvimento continuado. Termino com alguns comentários inspirados pelo pensamento de Mariano Gago sobre as relações entre a ciência e a tecnologia e a sociedade, explicando o porquê do título.

1. O RECURSO

O corpo Mariano Gago foi apresentado em Natal, no Brasil, em 2015 no encontro STIL (Santos, 2015b), por três razões: em primeiro lugar, por ser a conferência associada ao processamento computacional do português temporalmente mais próxima da sua criação (novembro de 2015); em segundo, porque aceitava artigos em português; e em terceiro, porque um dos cavalos de batalha de José Mariano Gago era precisamente o português como língua de peso a nível internacional, e daí os seus aturados esforços por uma cooperação com o Brasil nessa área específica[1].

Sendo Mariano Gago, por razões naturais, menos conhecido no Brasil do que em Portugal, também me pareceu importante que a sua fama se espalhasse e fosse claramente documentada em relação à Linguateca, que, essa, a maioria dos brasileiros da área do processamento de linguagem natural (PLN)

[1] A colaboração com o Brasil foi sempre uma constante ao longo da existência da Linguateca, e embora de forma alguma possa reclamar o apoio de Mariano Gago para as minhas ideias, visto que foram apenas desenvolvidas e fixadas em forma final depois da sua morte, embora inspiradas pelo meu trabalho na Linguateca e na minha vida profissionnal em geral, gostaria de referir a minha defesa de um português internacional em Santos (2016).

conheciam há muito. Mas não é comum que um projeto tenha a bênção, e a facilitação, de um político como Mariano Gago, e tal relação era importante que fosse conhecida também além-fronteiras[2].

Como este recurso foi apresentado numa conferência científica, exigia, além do testemunho de gratidão que foi o seu claro motor, uma mensagem de cariz técnico que o tornasse relevante por razões outras que históricas. Foi, portanto, apresentado, em primeiro lugar, como um recurso original – por ser o primeiro corpo linguístico recolhido em torno de uma personalidade, pensamos que a nível mundial, mas pelo menos em língua portuguesa; em segundo lugar, como um exemplo a seguir para criar outros corpos da mesma índole, e que pudesse servir para desenvolver ferramentas com um interesse para além da área técnico-científica, em particular contribuindo para a sociedade em que nos enquadramos e para cujo progresso trabalhamos; além de ser, claro, mais um recurso público a ser usado pela comunidade de processamento computacional do português.

Na foto documental temos o Robercy Alves, aluno de doutoramento em Natal na altura e monitor na conferência, à frente do cartaz apresentando o recurso.

É hora de o apresentar novamente.

FOTO DOCUMENTAL

1.1 O conteúdo deste corpo em particular

Se algum internauta incauto chegasse à página do projeto AC/DC onde se pode interrogar o corpo JMG[3], encontraria – desde 2015 – a seguinte descrição:

> "O corpus JMG contém artigos e notícias associadas a José Mariano Gago: obituários; testemunhos por ocasião da sua morte a 17 de abril de 2015; textos escritos por ele: discursos, cartas, outros textos; entrevistas; artigos sobre ele e a sua política; depoimentos de homenagem. Todos os textos em questão foram retirados da internete. Veja a página sobre o corpo para mais informação."

[2] A este respeito quero contudo salientar que, se Mariano Gago foi o iniciador e o padrinho, a Linguateca deve muitíssimo também ao apoio continuado de Pedro Veiga, que foi quem, durante mais de dez anos, recebendo ou não orientação de Mariano Gago (visto que anos houve em que este não estava no governo), manteve a existência da Linguateca sem sobressaltos, lidando com todos os assuntos burocráticos e administrativos e permitindo a existência concreta do projeto e de todos os seus intervenientes. Tal deve ficar registado, sobretudo porque, ao contrário de José Mariano Gago, Pedro Veiga não era especialmente fã da área, e penso que teria preferido um encaminhamento de fundos para tarefas e projetos mais do seu agrado.

[3] https://www.linguateca.pt/acesso/corpus.php?corpus=JMG

E se ficasse com mais curiosidade, poderia confirmar que existia uma outra página com mais informação[4], com última atualização em 9 de abril de 2016[5], assim como listas de fontes e outras informações técnicas, que uma pessoa realmente curiosa ainda poderia (e pode) ir consultar. Uma frase, contudo, poderia dar a entender ao nosso internauta que isto não era para ele:

> "A ideia é que esta coleção possa ser usada para treinar, testar ou validar sistemas que trabalham sobre a língua portuguesa, nas mais variadas formas. Este é portanto um trabalho em progresso, que esperemos que fomente o progresso das ferramentas criadas pela comunidade."

Que comunidade? Que ferramentas? Mas, se fosse realmente movido por um instinto detetivesco, poderia ir tentar usar a primeira página em que tinha caído, que tinha uma caixa de pesquisa (talvez?) apetecível. E poderia tentar procurar os exemplos que lhe são sugeridos, nomeadamente a forma *amigo*, frases contendo *Algarve*, o substantivo *político*, emoções, e nomes próprios contendo *Mariano*. Contudo, para saber usar as várias potencialidades, e poder considerar este sítio na internete ele próprio como uma ferramenta útil, o nosso internauta cedo compreenderia que teria de empregar muito mais tempo para se inteirar das possibilidades, da sintaxe da procura, etc. E provavelmente não voltaria, a não ser se fosse especialmente interessado em língua e em informática.

Confessamos, a interface não está virada para o público em geral. Somos investigadores, queremos serviços específicos para a "nossa" comunidade, não temos recursos (sobretudo humanos) para ensinar. Mas, como todos os cientistas, temos a esperança de um dia alguém nos descobrir, apostar em nós, basear os seus serviços (públicos ou comerciais) naquilo que iniciámos. Esta atitude não é suficientemente boa, e José Mariano Gago foi uma das pessoas que mais fez para a contrariar, interessado em que a ciência e a tecnologia chegassem ao povo e à sociedade.

Nesse campo a única atitude de que nos podemos orgulhar é não termos nunca criado mais barreiras para o uso dos nossos recursos. De facto, além de exemplificar o seu uso, tornamo-los públicos para que outros possam usá-los melhor do que nós. Afinal de contas, uma das ideias chave que preside à Linguateca é: "**Total abertura:** Todas as actividades e trabalhos desenvolvidos pela Linguateca são públicos[6]."

Neste artigo, vamos tentar que a leitora (não imediatamente uma internauta) compreenda melhor o que significa compilar um corpo de língua e que coisas se podem fazer com base no mesmo[7].

---

[4] https://www.linguateca.pt/CorpoJMG/

[5] Isto até ao dia 30 de março de 2018, altura em que foi atualizada.

[6] https://www.linguateca.pt/

[7] O uso da palavra *coisas* nesta frase é intencional. Não resisto a lembrar que uma das primeiras atividades de Mariano Gago em Portugal no campo da divulgação da ciência foi a exposição "De que são feitas as coisas?", e que depois disso fez muita, muitíssima coisa...

1.2 O que é um corpo linguístico e suas aplicações; o advento das Humanidades Digitais

Para as pessoas que estudam a língua – e que nessa função têm sido chamados gramáticos, filólogos, linguistas, especialistas de retórica, ou de literatura, engenheiros da linguagem, filósofos da linguagem, sociolinguistas, linguistas computacionais, lexicógrafos, lexicólogos, terminólogos, fonologistas, dialetologistas, etc., etc., dependendo da época e da área em que melhor se enquadram –, o seu objeto de estudo tem variado conforme as perguntas relevantes e as tecnologias acessíveis, e conforme a atitude do investigador em relação ao objeto de estudo.

Um corpo linguístico em formato digital é apenas uma forma de obter material de estudo para compreender a língua, forma que é apoiada pela tecnologia informática existente. Nem todos os que estudam ou estudaram a língua precisam ou precisaram de um corpo, mas estamos convencidos de que cada vez mais estes serão imprescindíveis.

A língua tem várias vertentes e em particular tem duas formas físicas de existir: sonora, e escrita, a última das quais evidentemente muito posterior, mas inventada precisamente para desafiar o tempo e guardar o que se diz ou sabe, mesmo depois de quem o disse ou sabia ter desaparecido. Com o advento da tecnologia de gravação e, depois, da computação, encontramo-nos numa situação em que é igualmente possível guardar em formato digital a fala e a escrita. Contudo, de um ponto de vista filosófico, as duas são muito diferentes – enquanto uma é imediata, a outra é redigível e alterável consistindo numa série de versões; uma é dirigida para a interação, outra para a informação; enquanto a primeira é volátil, a segunda ainda tem o peso de "ser escrita", ser pensada, ter sido publicada, ter passado por vários crivos de qualidade. Até que ponto essa dicotomia continuará na nossa sociedade após o nascimento de vários tipos de "falar digital" como os pios (*tweets*), os bate-papos (*chat rooms*), os sms e os comentários dos leitores nos blogues e nas redes sociais, é algo sobre o qual muitos especialistas dos meios de comunicação social, assim como sociólogos, especulam no momento presente. Eu apenas menciono que, nascidos digitais, ainda mais óbvio é que existem corpos de todas estas formas de comunicação, assim como de filmes ou vídeos.

Mas o que é, afinal, um corpo de língua? Por uma questão de simplicidade, vou-me ater apenas aos corpos que estão num formato escrito – mas repare-se que transcrição de discursos, de debates na Assembleia da República, ou de entrevistas, é algo que existe há muito (muito antes, aliás, de se usarem corpos). E todos sabemos que grandes obras literárias como as de Homero ou os contos populares foram forjados e preservados pela transmissão oral, muito antes de terem sido "passados para o papel". A transcrição, contudo, é importante salientar, é sempre uma interpretação, e não algo objetivo. Exige sempre escolhas e cria novos textos.

Um corpo de língua é, então, um conjunto de textos, escolhidos com um dado fim, e anotados em relação a esses critérios de escolha, para funcionar em geral como uma amostra da língua que se quer estudar, veja-se Santos (2008). Na maior parte dos casos, além dessa anotação (frequentemente referida como metadados), os corpos também são anotados com informação linguística relacionada com o seu

próprio conteúdo. Por exemplo, expressões idiomáticas podem estar marcadas como tal, ou a categoria gramatical de cada palavra, ou simplesmente os casos de erros (e sua correção), ou os estrangeirismos, ou o tempo verbal, ou as funções sintácticas (sujeito, objeto direto, predicativo do sujeto, etc.), ou mesmo o tema (por exemplo: peça de roupa, parte do corpo, emoção positiva, cor, etc.). Em princípio, um corpo pode ser anotado com qualquer informação em que uma investigadora esteja interessada, exatamente para depois poder estudar de uma forma quantitativa o resultado.

É importante salientar que anotar um corpo não é, principalmente, uma forma de preparar dados para realizar mais tarde estudos quantitativos. De facto, a necessidade de estabelecer regras descritivamente válidas para qualquer assunto é um dos desafios mais interessantes da linguística com corpos – que, por oposição a uma linguística teórica, olha para a língua como é usada para chegar a conclusões sobre a sua constituição e o seu funcionamento.

Nesse aspeto, podemos dizer que a linguística com corpos é uma das áreas da ciência experimental (por analogia à física experimental) e portanto uma base (e não uma competidora) em relação à linguística teórica. Contudo, como é comum na academia, existe, tal como na física, alguma fricção entre estes dois campos, embora uma outra dicotomia esteja cada vez mais na ordem do dia, relacionada com a utilização de métodos computacionais, em particular estatísticos, para as grandes tarefas da linguística computacional, em detrimento da análise humana.

E isto leva-me à nova moda (não necessariamente no mau sentido) das Humanidades Digitais, que é algo que, pelo menos na área mais ligada à língua e em particular à literatura, defende precisamente o uso de técnicas computacionais "incompreensíveis aos seres humanos" para chegar a novas perspetivas e conhecimento sobre áreas diretamente relacionadas com a atividade humana. Inspirada pela física, pela química e pela astronomia, uma das ideias básicas dos métodos digitais para o estudo da literatura é o conceito de macroscópio, um aparelho (metafórico) que permite ver um assunto a distância, tal como foi possível fotografar a Terra pela primeira vez do espaço. Mais do que isso, permite escolher a distância a que se "visualiza", e observar o campo de estudo de várias perspetivas diferentes. Perspetivas essas, contudo, que apenas um computador (ou melhor, um sistema computacional) consegue "apreciar", visto que, por exemplo, consegue "ler" milhões de livros/obras, algo que mesmo o mais interessado especialista de literatura não conseguiria. E, por isso, a quantidade torna-se apenas "legível" através de máquinas poderosas, que usam técnicas estatísticas para, por exemplo, reduzir a literatura a um conjunto de eixos interessantes.

Tal não é obviamente nada de novo, dado que os motores de procura na internete usam as mesmas técnicas (ou técnicas semelhantes) para procurar o que poderíamos apelidar de verdadeiras agulhas em palheiro. Mas traz algumas questões que nos deviam levar a uma maior reflexão. Nomeadamente, como distinguir entre resultados automáticos, e outros que lá são postos maliciosamente? Como garantir que as técnicas usadas não têm vieses que sejam perigosos – como garantir, em última análise, que a sua avaliação é bem feita? Ao contrário do que as pessoas ouvem sobre a aprendizagem automática (*machine learning*) como sendo a resposta a todos os desafios científicos (veja-se o livro *Homo Deus* (Harari,

2015) para uma excelente denúncia desta postura), esta é apenas um conjunto de técnicas estatísticas com muitíssimos graus de liberdade, no sentido de que há muito ainda a aprender e muitas escolhas a fazer. Ora, é essencial poder avaliar o que é que essas técnicas conseguem e não conseguem antes (ou em vez) de as endeusar. Muitas vezes, contudo, as avaliações são precisamente viradas para o desempenho que uma dada técnica consegue: nesse caso, são mais prejudiciais do que benéficas.

Voltando à questão dos corpos de língua, e falando agora da aplicação dessas técnicas à língua, considero que a existência de corpos especificamente criados e controlados em relação a um dado assunto ou tarefa é uma espécie de antídoto contra a confiança cega em técnicas automáticas desenvolvidas por companhias privadas.

Certamente o seu tamanho é de muitas ordens de grandeza abaixo do que as coleções (ou corpos) que companhias como a Google ou a Facebook controlam, mas são públicas, são repetíveis, e são passíveis de ser discutidas e criticadas – de serem, numa palavra, objeto de ciência. E, além disso, no caso do português, são uma garantia de que a nossa língua não é completamente engolida e triturada por interesses de multinacionais.

1.3 O papel da Linguateca em relação aos corpos do português

É aqui portanto que entra a criação da Linguateca, que planeio, aliás, descrever em maior detalhe em breve[8], cujas sementes surgiram durante o processo de organização da discussão pública sobre o processamento computacional da língua portuguesa, que teve lugar em abril de 1999, e cuja razão de ser foi melhorar as condições para se fazer ciência na área da língua portuguesa, considerando o português em todo o mundo e não só o de Portugal.

Daí nasceu o projeto AC/DC, que foi ganhando momento como uma bola de neve. Iniciado em 1999, esta iniciativa foi congregando mais recursos e mais ferramentas, servindo mais tarde de base para outras infraestruturas: a Gramateca, reunindo interessados em gramática baseada em corpos, e a Literateca, estimulando a leitura à distância das obras literárias lusófonas. Embora o AC/DC tenha sido por várias vezes descrito[9], é preciso indicar que a colaboração com Eckhard Bick com a consequente introdução da análise sintática do PALAVRAS na linha de montagem dos corpos foi decisiva para a qualidade deste serviço, que é hoje ensinado a alunos de português como cadeira universitária, e sobre o qual tem havido várias ações de formação em Portugal e no Brasil.

---

[8] Entretanto, podem ser consultados, para os que gostam de História, textos como Santos (2000), Costa et al. (2008), Santos (2009), Santos (2015a) e o prefácio de Santos (2007).

[9] Veja-se Santos & Ranchchod (1999), Santos & Bick (2000), Santos & Sarmento (2002), Costa et al. (2007), Santos (2011) e Santos (2014).

Por limitações de espaço, e por tudo isto já ter sido documentado nos textos acima mencionados, a única coisa que me parece fazer sentido dizer neste contexto é que existem dois tipos de corpos servidos pelo AC/DC: aqueles em que o acesso é apenas feito através de pesquisas (por razões de direitos de autor, ou direitos dos compiladores dos corpos) e aqueles que também são disponibilizados em texto completo e com anotações, como é o caso do corpo Mariano Gago.

Consideramos que é útil oferecermos estas duas formas de disponibilização, porque temos dois tipos de utilizadores distintos, os linguistas e os engenheiros da linguagem. Se, para os engenheiros, a opção preferida é sempre a obtenção do material completo, para sobre ele testarem ou aplicarem as suas ferramentas, para muitos linguistas a possibilidade de simplesmente usarem a internete para fazer a sua investigação foi (e ainda é) revolucionária. Por outro lado, a maior quantidade de material e o contínuo progresso na anotação fazem com que as versões acessíveis através do serviço na internete sejam em geral mais avançadas e mais ricas do que as prontas para levantamento.

2. O QUE SE PODE FAZER COM ESTE RECURSO

Mas vejamos o que já se pode fazer com este recurso: exemplos concretos de procuras e de interrogações a que o corpo poderia responder – e exemplos concretos de tecnologias que podem ser desenvolvidas com ele ou sobre ele aplicadas. Em seguida, especularemos sobre o futuro se viéssemos a ter acesso a muito mais material e a dispor de trabalho humano de revisão de anotação.

2.1 O que já se pode fazer

As duas tarefas básicas que todos os corpos linguísticos oferecem, são as concordâncias, e as distribuições. As concordâncias mostram aquilo que se pesquisou em contexto (veja-se a Figura 1), e as distribuições mostram, como o nome indica, como aquilo que pesquisámos se distribui segundo uma característica específica (o género textual, o tempo verbal, a forma canónica, o autor, etc. etc.).

No momento presente, a leitora impaciente pode já obter várias ideias sobre o conteúdo deste corpo, simplesmente fazendo perguntas sobre que nomes próprios lá aparecem[10]. (Ou seja, pedindo a distribuição dos nomes próprios pelo seu lema.) E assim inteirar-se das variadas instituições ou organizações associadas a Mariano Gago, como por exemplo a *Agência de Avaliação e Acreditação do Ensino Superior*, ou o *Laboratório de Física Nuclear e de Altas Tecnologias da École Polytechnique*. Ou também poderia pesquisar quais os verbos mais empregues no material.

[10] Criámos uma página com os exemplos detalhados em https://www.linguateca.pt/CorpoJMG/ExemplosCorpoJMG.html

Mas dado que o conteúdo das notícias após a morte é certamente diferente do das notícias anteriores, poderá ser interessante contrastar exatamente isso: e com efeito, *morrer* e *deixar* são dos mais frequentes no primeiro grupo, contrastando com verbos como *haver, dever* e *ir*, que são mais usados no segundo.

Também se pode investigar a proeminência ou ausência de campos semânticos como a roupa, o corpo ou a cor. Nenhum deles muito interessante à partida, mas basta começar a interrogar para, por exemplo no caso da roupa, Mariano Gago ser lembrado pela "sua simples gabardina", a "canadiana escura" (nos tempos de estudante) e o seu "entrar em mangas de camisa", ou seja, a sua informalidade.

Mas claramente os campos mais interessantes neste recurso são o das emoções e da fala. Se escolhermos, por exemplo, o subconjunto proveniente do sítio de homenagem, as (palavras associadas a) emoções ali referidas são, por ordem decrescente, *gostar, querer, saudade, acreditar, reconhecimento, luto* e *pesar*. Se olharmos para as notícias após a sua morte, teremos *pesar, querer, reconhecimento* e *lamentar*. Veja-se a Figura 2.

FIGURA 1

Resultados da procura

5 de abril de 2018

Procura: **MU (meet [lema="cultural?"] [lema="ciência|científico"] s)**
Pedido de uma concordância em contexto
Corpo: Textos de ou sobre José Mariano Gago v. 3.1

256 ocorrências.

Concordância

Procura: **MU (meet [lema="cultural?"] [lema="ciência|científico"] s)**.

t1-19: «Inestimável contributo para a ciência, tecnologia e a **cultura** científica em Portugal» Fundação para a Ciência e Tecnologia

t1-20: A Fundação para a Ciência e Tecnologia (FCT) lamentou esta sexta-feira, com «profundo pesar», a morte do antigo ministro Mariano Gago, destacando o seu «inestimável contributo para a ciência, tecnologia e a **cultura** científica em Portugal» .

t1-26: O Presidente da República, Aníbal Cavaco Silva, enviou, esta sexta-feira uma mensagem de condolências à família do professor José Mariano Gago, que morreu, frisando que Portugal perde «uma das personalidades mais marcantes da sua vida científica e **cultural**» .

t1-27: «Além de homem de ciência e **cultura**, José Mariano Gago foi, desde jovem, um cidadão exemplar pelo empenho demonstrado na defesa intransigente dos valores da liberdade e da democracia», frisa a nota .

t4-8: Inspirador, sonhador, energético, líder, alguém que teve uma visão para a ciência e a **cultura** científica do país e, mais do que isso, a pôs em prática foram palavras usadas na sexta-feira para o descrever .

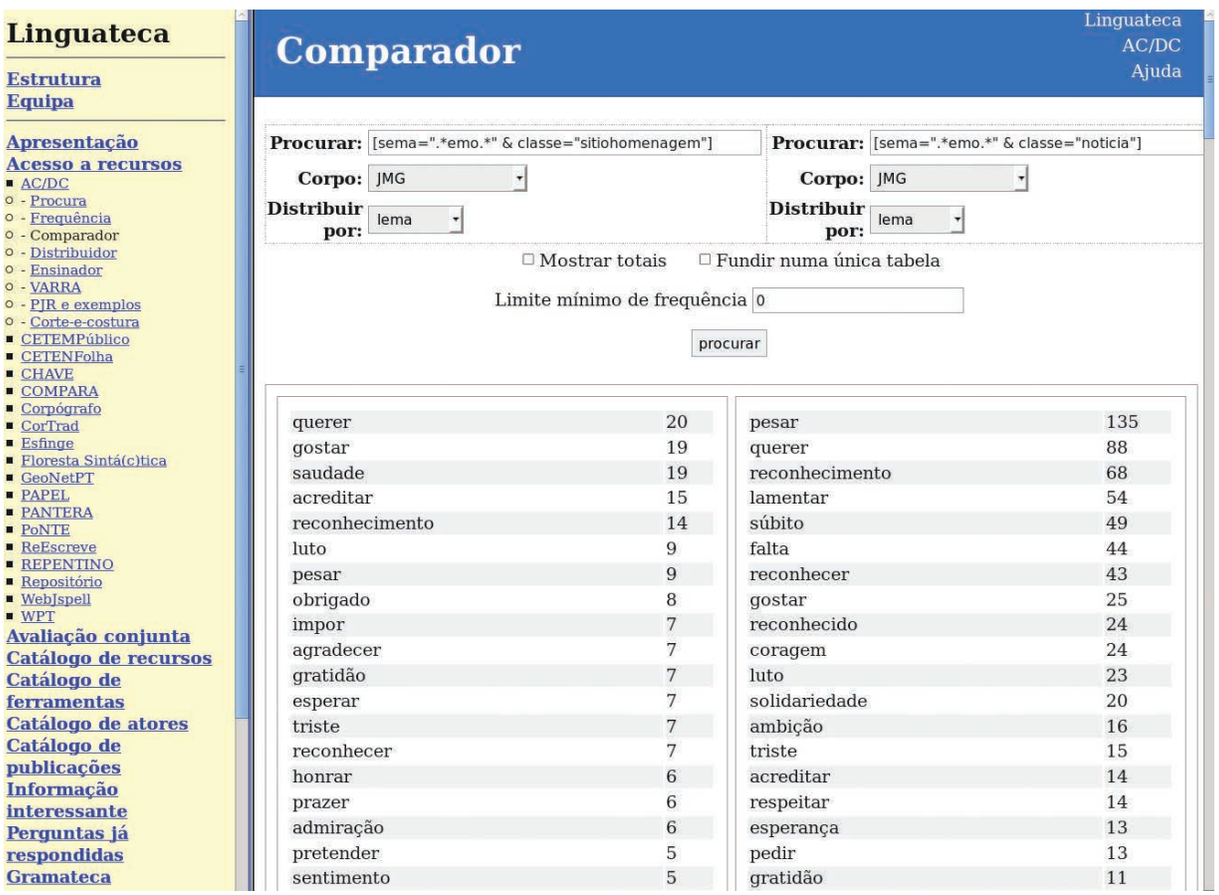
t4-17: Muitos cientistas estrangeiros perguntam-nos: Vocês têm uma agência só para a **cultura** científica ?

t7-7: José Mariano Gago «colocou no centro da ambição política a sociedade do conhecimento e teve uma preocupação incansável com a democratização da **cultura** científica, designadamente através do lançamento do programa Ciência Viva .

t16-23: Costa falou ainda da «preocupação incansável pela democratização da **cultura** científica», protagonizada por Mariano Gago, considerando o lançamento do programa ciência viva .

t22-7: Além de homem de ciência e **cultura**, José Mariano Gago foi, desde jovem, um cidadão exemplar pelo empenho demonstrado na defesa intransigente dos valores da liberdade e da democracia .

FIGURA 2



Pode ser controversa a inclusão de *querer* no campo das emoções mas, independentemente dessa escolha, é interessante observar os usos do verbo *querer*: por um lado, descrições de intenções; por outro, modalizadores: *Quero afirmar, quero prestar homenagem, não quero deixar de, quer assinalar, ...* Os primeiros dão uma ideia clara de um homem que quis e fez, enquanto os segundos correspondem a um coro de homenagem.

Se investigarmos o campo do discurso relatado (o único que foi até agora trabalhado ativamente com base neste corpo, tendo os resultados sido apresentados no LREC[11] de 2016), pode ser interessante identificar algumas afirmações e alguns relatos presentes no material:

[11] LREC *(Language Resources and Evaluation Conference)* é a conferência internacional mais importante no campo dos recursos e da avaliação, e o artigo referido é Freitas *et al*. (2016).

Ficamos a saber de uma frase bonita, e quem foi o seu autor:

Como **disse** *o Miguel Esteves Cardoso: «Viveu Mariano Gago»*

E, se ficássemos interessados em (re)ler esse depoimento, e procurássemos então todas as formas do verbo *viver*, obteríamos mais algumas citações interessantes, como

Mariano Gago ***viverá*** *enquanto a Ciência viva*
Morreu o homem que ***vivia*** *a ciência em Portugal*

E seríamos também lembrados de um percalço provocado pelas condolências pouco elegantemente formuladas pelo primeiro-ministro da altura, que levaram a muitos comentários, dos quais seleciono apenas uma pequena amostra:

Passos Coelho, Mariano Gago e a gafe inoportuna
Como dizia um amigo meu, Pedro Passos Coelho inaugurou um novo voto, o voto de apesar.

[...] o modo desastrado como Passos Coelho se referiu ao acontecimento [...]

A descabida reserva política («apesar de ter sido ministro socialista... ») colocada por Passos Coelho na displicente homenagem fúnebre ao notável cientista e universitário que JMG foi não revela somente sectarismo político mas também falta de chá democrático .

[...] a afirmação de Passos Coelho reflecte «uma falta de sensibilidade chocante», [...]

[...] muito triste episódio do «elogio» fúnebre do Passos ao Mariano Gago [...]

Embora certamente o mais interessante será de facto ler, por exemplo, o que o próprio Mariano Gago disse na entrevista dada a Luísa Tiago de Oliveira, ou nos seus discursos – era um orador como poucos, provavelmente dadas as suas incursões no teatro enquanto estudante.

Mas as vozes dissonantes ou as afirmações contra a sua atividade também são de manter e de considerar, e é interessante observar que também existem, neste recurso, e na sua vida. Tenho a certeza de que ele gostaria que o material fosse tão fidedigno quanto possível, para se poder também estudar a conjuntura e as reações às suas medidas ou posições, assim como o papel dos meios de comunicação social na cobertura da (política da) ciência em Portugal.

Nesse aspeto convém realçar que um estudo superficial da sua presença no jornal Público, usando o corpo CETEMPúblico (de 1991 a 1998), é francamente positivo no que se refere às medidas concretas, mas encontra-se quase exclusivamente nas seções de cultura e sociedade. Apenas 20 citações provêm de uma secção de opinião, todas de 96 e de 97, e por três vezes a falta de aparência mediática é posta em destaque. Gostaria de as citar aqui, pelo flagrante contraste em relação a 2015:

A primeira, do primeiro semestre de 1997, é elucidativa: *Até agora, o nome do ministro Mariano Gago só aparecia nos jornais por sarcasmo, dado como «ausente».* A segunda, de julho de 1996, reza

assim: *No canto inferior esquerdo da página 10 do Público de 6 de Julho, encontra-se um rectângulo de cerca de cem palavras a propósito do ministro da Ciência Mariano Gago, o tal que ninguém conhece, aquele cujo trabalho as pessoas em geral (e os jornalistas) desconhecem.* E a terceira proclama no seu título: *Mariano Gago É um ministro que não existe, a julgar pela sondagem efectuada pela Universidade Católica para o Público.*

A comunicação social foi certamente obrigada a rever a sua impressão e render-se ao facto de que Mariano Gago foi provavelmente o ministro mais querido de todos os portugueses!

Mas não nos afastemos demasiado do nosso recurso, porque a segunda vertente do seu desenvolvimento praticamente ainda não foi mencionada aqui, e é igualmente relevante (embora não possa ser testada pela nossa leitora, que, esperamos, já se tenha tornado uma futura utilizadora do recurso como fonte de informação). O segundo objetivo deste corpo era ser usado como semente (ou matéria-prima) no desenvolvimento de uma série de programas com objetivos concretos e práticos e, mais tarde, para avaliação dos mesmos. Resumindo o cartaz apresentado em Natal, a saber: Desenho de linhas temporais; criação e teste de sistemas mais complexos de reconhecimento de entidades mencionadas; identificação de discurso relatado; criação de material didático; estudos sobre emoções; estudos sobre relações sociais: quem refere quem, quem se relaciona com quem?

Outros programas e tarefas para cujo teste este corpo é naturalmente apropriado, mencionados no artigo embora sem lugar no cartaz, são: o estudo do comportamento dos meios de comunicação social na internete (frequência de atualização de notícias, da escrita de novas notícias, o reuso de outros textos); a criação de perguntas de compreensão para o ensino do português como língua estrangeira; a análise de reputação; a classificação automática de géneros textuais; a limpeza de textos repetidos; a identificação de fontes (explícitas e implícitas) de uma notícia; e, claro, a sumarização inteligente, com a correspondente visualização.

Enquanto algumas destas áreas ainda constituem investigação de ponta, outras há que se encontram já razoavelmente bem desenvolvidas, mas para todas a existência de um conjunto calibrado e analisado por seres humanos é uma mais-valia considerável, e leva-nos a uma das áreas chave da atividade da Linguateca, área que Mariano Gago apreciava especialmente: a da avaliação em ciência.

2.2 Potencialidades futuras

Agora, é este recurso suficiente para documentar a presença de Mariano Gago na ciência e tecnologia portuguesas, ou sequer apenas na área do processamento computacional da língua? Claro que não. Muito mais teria de ser trazido à colação, e aproveito esta oportunidade para o referir.

Se continuássemos a desenvolvê-lo, uma ação natural, que exigiria colaboração com outros atores, neste caso obviamente o Arquivo da Web, seria estudar, de forma diacrónica, a presença de Mariano Gago na internete. Outra seria, em colaboração com os maiores jornais portugueses (e canais de rádio e televisão), levar a cabo o mesmo estudo em cada um.

O conjunto de discursos e de escritos político-científicos de Mariano Gago, o seu espólio, foi doado à FCT. A possibilidade de ter acesso aos mesmos em forma digital seria outra maneira de aumentar o presente recurso.

Ainda melhor seria transformá-lo num ponto de acesso à obra e ao legado dele, enriquecido pela tecnologia do processamento de linguagem natural, do português, para o qual ele tanto contribuiu e no qual tanta esperança tinha. Em vez de um simples corpo ou de um arquivo tradicional, poderíamos criar um "Mariano Gago virtual" que seria construído, e mantido, em prol da ciência e do processamento da língua portuguesa.

3. OS ENSINAMENTOS DE JOSÉ MARIANO GAGO

Houve muitas coisas que José Mariano Gago me ensinou, pelas quais estou profundamente grata. Cabe mencionar duas a respeito do presente projeto: a primeira, a urgência de levar a ciência, e o conhecimento, para fora da universidade, para a sociedade e para a política, como já por várias vezes aludi. E a segunda, a necessidade de documentar o que foi feito, para permitir às gerações vindouras evitar erros, aprender com o passado e construir sobre o que já existe.

Este texto dá uma pequena achega para a documentação cabal de tudo o que Mariano Gago fez e representou. O seu foco, o corpo Mariano Gago, foi uma tentativa de criar algo que pudesse servir a pessoas com muito mais perfis do que apenas engenheiros da língua ou linguistas computacionais: profissionais da comunicação social, historiadores, simples curiosos...

Se conseguirmos obter utilizadores com outros perfis, e isso possa contribuir para um maior conhecimento público de José Mariano Gago e do seu legado, o nosso esforço não terá sido em vão. É, evidentemente, da mais elementar justiça que tenha sido a Linguateca a produzir, ou iniciar, este recurso. Se, além disso, conseguirmos entusiasmar futuros colaboradores para tornarem este projeto muito maior, então conseguiremos ultrapassar as fronteiras da academia e ter utilidade sócio-cultural, como o nosso amado "ministro da língua" sonhava para a ciência em geral.

AGRADECIMENTOS

Embora o artigo tenha sido escrito em meu nome, não posso deixar de mencionar todos os meus colegas na Linguateca que continuam a trabalhar e a desenvolver o campo do processamento computacional da língua portuguesa porque nele acreditam, e sem os quais a Linguateca não teria continuado: em particular, nos últimos tempos, quero agradecer especialmente ao Alberto Simões, à Cláudia Freitas, à Cristina Mota e ao Luís Miguel Cabral.

REFERÊNCIAS

Costa, Luís, Diana Santos & Nuno Cardoso (eds.). *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos.* Linguateca. 2008. http://www.linguateca.pt/LivroL10/

Costa, Luís, Diana Santos & Paulo Alexandre Rocha. "Estudando o português tal como é usado: o serviço AC/DC". *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)* (São Carlos, Brasil, 8-11 de Setembro de 2009).

Freitas, Cláudia, Bianca Freitas & Diana Santos. "QUEMDISSE?: Reported speech in Portuguese". In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* pp. 4410-4416.

Harari, Yuval Noah. *Homo Deus: A brief history of tomorrow.* Penguin, 2016. (Versão em hebraico, 2015).

Santos, Diana. "O projecto Processamento Computacional do Português: Balanço e perspectivas". In Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)* (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), pp. 105-113.

Santos, Diana (ed.). *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa.* IST Press, 2007.

Santos, Diana. "Corporizando algumas questões". In Stella E. O. Tagnin & Oto Araújo Vale (orgs.), *Avanços da Lingüística de Corpus no Brasil,* Editora Humanitas/FFLCH/USP, São Paulo, 2008, pp. 41-66.

Santos, Diana. "Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva". *Linguamática* 1, 1, Maio de 2009, pp. 25-58.

Santos, Diana. "Linguateca's infrastructure for Portuguese and how it allows the detailed study of language varieties". In J. B. Johannessen (ed.), *Language variation Infrastructure, OSLa 3 (2),* 2011, pp. 113-128.

Santos, Diana. "Corpora at Linguateca: Vision and roads taken". In Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora,* Bloomsbury, 2014, pp. 219-236.

Santos, Diana. "Portuguese language identity in the world: adventures and misadventures of an international language". In Elizaveta Khachaturyan (ed.), *Language - Nation - Identity: The questione della lingua in an Italian and non-Italian context,* Cambridge Scholars Publishing, 2015a, pp. 31-54.

Santos, Diana. "Um novo corpo e seus desafios". In Cláudia Freitas & Alexandre Rademaker (eds.), *STIL 2015: X Brazilian Symposium in Information and Human Language Technology and Collocated Events, Proceedings of the Conference. November 4 to 7, 2015.* Natal, Rio Grande do Norte, pp. 39-43.

Santos, Diana. "Português internacional: alguns argumentos". In José Teixeira (ed.), *O Português como Língua num Mundo Global: problemas e potencialidades,* Centro de Estudos Lusíadas da Universidade do Minho, 2016, pp. 49-66.

Santos, Diana & Eckhard Bick. "Providing Internet access to Portuguese corpora: the AC/DC project". In Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000 (Athens, 31 May-2 June 2000),* pp.205-210.

Santos, Diana & Elisabete Ranchhod. "Ambientes de processamento de corpora em português: Comparação entre dois sistemas". In *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR* (Évora, 20-21 de Setembro 1999), pp. 257-268.

Santos, Diana & Luís Sarmento. "O projecto AC/DC: acesso a corpora / disponibilização de corpora". In Amália Mendes & Tiago Freitas (orgs.), *Actas do XVIII Encontro da Associação Portuguesa de Linguística* (Porto, 2-4 de Outubro de 2002), APL, 2003, pp. 705-717.

Ficha Técnica

Título:
Caminhos do Conhecimento, o Legado de José Mariano Gago.
Dia Nacional dos Cientistas.

Coordenação:
Rosalia Vargas, Ana Noronha, Carlos Catalão Alves

Edição:
Carlos Catalão Alves

Design:
Marisa Vinha

Revisão:
Domingas Portela, Gonçalo Praça

Composição, impressão e acabamento:
Empresa Diário do Porto

Tiragem:
1ª edição - Abril de 2018
500 exemplares

Depósito Legal:
425339/17

ISBN:
978-972-98602-5-6

Ciência Viva
Agência Nacional para a Cultura Científica e Tecnológica
Pavilhão do Conhecimento,
Largo José Mariano Gago, 1
Parque das Nações
1990-223 Lisboa

Os textos respeitam a liberdade de escolha dos autores quanto à ortografia usada.