Apresentando os recursos da Linguateca

Diana Santos

d.s.m.santos@ilos.uio.no

Escola Superior de Educação de Bragança, 21 de junho de 2019

Apresentação

O que me foi pedido

> *sensibilizar os alunos da ESEB para a importância do uso rigoroso da Língua Portuguesa nas suas vertentes oral e escrita*

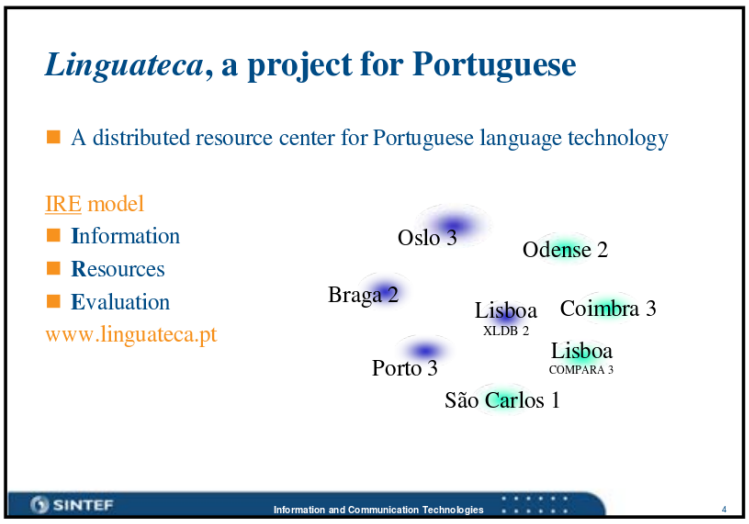
O que é a Linguateca

A Linguateca foi um projeto pensado (em 1999) para os engenheiros da linguagem, ou linguistas computacionais:

- para desenvolver recursos para serem usados na investigação e nos produtos
- para desenvolver ferramentas que fossem pensadas de raiz para a nossa língua, e não meras adaptações
- para avaliar, em conjunto, o progresso em várias sub-áreas
- para mentalizar no sentido da disponibilização livre e partilha de esforços

Não foi nunca um projeto para servir o grande público.

No seu auge, a Linguateca congregou



- vários pólos em diversos centros de investigação
- mais de 20 colaboradores

A partir de 2011, a Linguateca não mais teve financiamento. Mas temos alojamento informático nos servidores da FCCN.

Na última década

- virámo-nos mais para o ensino e pesquisa da própria língua (e da literatura)
- deixámos de poder catalogar o que se passava na área
- temos uma atividade muitíssimo menor
- mas tentamos que os recursos continuem a ser utilizados

A nossa língua

- Falada nas cinco partes do mundo
- Com uma história e uma literatura riquíssimas
- Com uma grande variedade que a enriquece

Mas: vai resistir à globalização tecnológica?

Várias questões, muito brevemente

- A importância da língua materna
- A importância da compreensão do modelo do mundo (dos modelos do mundo) associados a uma língua e cultura
- A importância da nossa consciencialização como falantes
- As atitudes perante a língua
  - As questões políticas associadas a investir ou não numa língua, e a preparação dos professores
  - As questões económicas subjacentes

Mariano Gago

A questão tecnológica

O sonho de todas as empresas globais é

- reduzir as línguas a simples escolha de um léxico diferente para dizer "a mesma coisa", ou um sotaque / entoação diferente para "fazer as mesmas tarefas"
- Primeiro fazem para o inglês (ou para o chinês) e depois adaptam para outras línguas, por ordem de importância económica (mercado).

Exemplos

Sempre que algo tem a palavra "universal", é de desconfiar...

- Universal dependencies – todas as línguas têm de ter as mesmas relações, e depois o problema é meter numa camisa de forças as particularidades de cada língua
- NRC Emotion lexicon – Emoções "universais" traduzidas do inglês para várias línguas

Estes são dois projetos que tratam de semântica e que mostram o que está em voga.

Ideias centrais da Linguateca

- Deve partir-se da própria língua, e não de modelos desenvolvidos para outras línguas
- Devemos partilhar ideias, hipóteses, e métodos, mas não aplicá-los cegamente
- O processamento de uma língua pode alterar a própria língua: algo que é evidente com o Google translate
- A comunidade que conhece a língua deve estudá-la e criar ferramentas próprias, e não simplesmente produzir dados.

Que recursos para esta audiência?

- A maior parte dos recursos que a Linguateca produziu ou mesmo catalogou são para o processamento computacional da língua. Ou seja, para serem usados por programas.
- Nesta apresentação, vou simplesmente dedicar-me a corpos e a serviços da rede que sejam para o uso humano.

Mas, mais importante do que os próprios recursos, é a consciencialização de que não é só traduzir, e de que cada língua tem um conjunto de características únicas que é interessante estudar e conhecer.

Audiência imaginada

- Línguas para Relações Internacionais
- Relações lusófonas e língua portuguesa
- Língua e cultura portuguesas
- Mestrado em tradução
- Educação básica

Convicções:

- Ter um bom domínio da sua língua e a consciência da importância desta são fatores importantes para a realização pessoal de um ser humano.
- A sensibilidade linguística para as diferentes nuances e consequências das escolhas na comunicação é recomendável para todas as profissões que lidam com pessoas.

Corpos linguísticos

> *Um corpo é uma coleção* **classificada** *de* **objectos linguísticos** *para* **uso** *em PLN/LC/L. Esses objetos são em geral também analisados/anotados/etiquetados.*

É portanto necessário ter a noção de quais os parâmetros de recolha, qual o universo. E qual o objetivo de criar essa coleção. Na prática, muitos corpos são criados oportunisticamente, e são recauchutados para novos usos.

As principais vantagens dos corpos

- Acervo de documentos autênticos, que abrange muito mais falantes do que nós próprios ou os nossos conhecidos
- A possibilidade de distribuição
- Dar azo à confirmação por outros das nossas descobertas ou afirmações
- Não ter limites de acessibilidade por questões de delicadeza: o computador está sempre lá para nos servir

Corpos na Linguateca

- CETEMPúblico
- Textos literários
- Jornais locais, e políticos
- Textos de divulgação científica (Ciência Viva)
- Textos de entrevistas a pessoas comuns (Museu da Pessoa)
- Textos relacionados com José Mariano Gago
- ... e muitos outros

Corpos paralelos

- COMPARA
- CorTrad
- PoNTE
- PANTERA

Pondo duas línguas em confronto, através da tradução.

O que é a língua portuguesa?

A única língua que tem um poema dedicado? A flor do Lácio, de Olavo Bilac... José Eduardo Agualusa, *O paraíso e outros infernos* (pag. 55f):

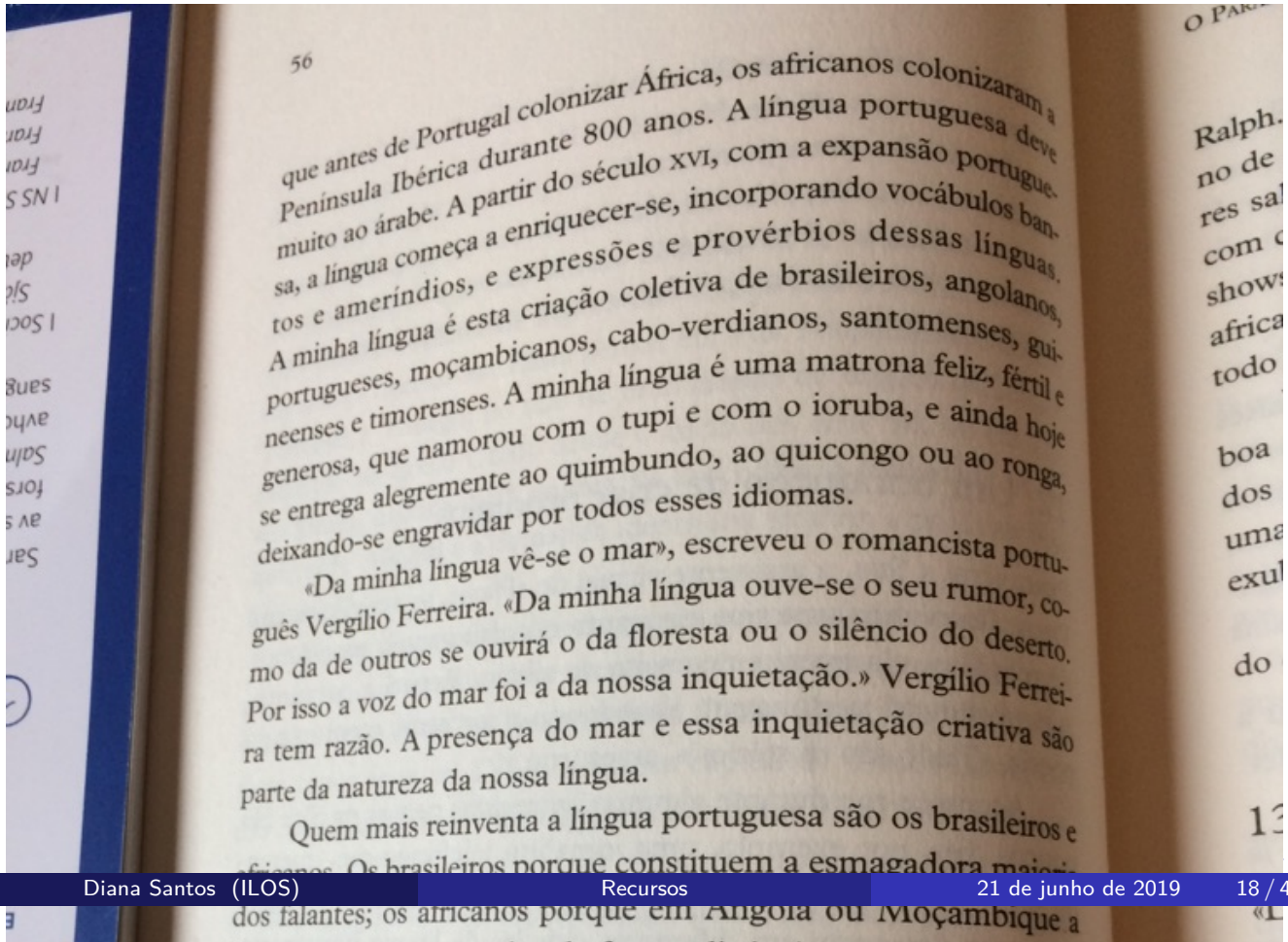
poucos dias, por exemplo, uma jornalista [illegible]
do saber como eu classificaria a língua em que escrevo: «Os seus romances decorrem em diferentes cidades de língua portuguesa, Luanda, Rio de Janeiro, Lisboa, até mesmo em Pangim, a capital de Goa (na Índia). Afinal, que língua portuguesa é a sua?»

Que língua portuguesa é a minha?!

Pensei em responder ao estilo da Siri: «Tudo são mistérios!» Infelizmente faltou-me a coragem e tropecei na resposta. Contudo, fiquei a pensar naquilo. Algumas coisas eu sei. Sei, desde logo, que a minha língua não está limitada por fronteiras políticas ou geográficas. O português que me interessa é o português total.

Há alguns anos, em Lisboa, num evento em que se discutia pela milésima vez o Acordo Ortográfico, um sujeito ergueu-se aos berros, no fundo da sala: «A língua é nossa!» Não fiquei surpreendido. A verdade é que ainda persiste em Portugal uma certa sau- [illegible] enorme ignorância no que diz respeito à história do próprio idioma. É sempre bom recordar

O que é a língua portuguesa?



56

que antes de Portugal colonizar África, os africanos colonizaram a Península Ibérica durante 800 anos. A língua portuguesa deve muito ao árabe. A partir do século XVI, com a expansão portuguesa, a língua começa a enriquecer-se, incorporando vocábulos bantos e ameríndios, e expressões e provérbios dessas línguas. A minha língua é esta criação coletiva de brasileiros, angolanos, portugueses, moçambicanos, cabo-verdianos, santomenses, guineenses e timorenses. A minha língua é uma matrona feliz, fértil e generosa, que namorou com o tupi e com o ioruba, e ainda hoje se entrega alegremente ao quimbundo, ao quicongo ou ao ronga, deixando-se engravidar por todos esses idiomas.

«Da minha língua vê-se o mar», escreveu o romancista português Vergílio Ferreira. «Da minha língua ouve-se o seu rumor, como da de outros se ouvirá o da floresta ou o silêncio do deserto. Por isso a voz do mar foi a da nossa inquietação.» Vergílio Ferreira tem razão. A presença do mar e essa inquietação criativa são parte da natureza da nossa língua.

Quem mais reinventa a língua portuguesa são os brasileiros e [illegible] Os brasileiros porque constituem a esmagadora maioria dos falantes; os africanos porque em Angola ou Moçambique a

O CETEMPúblico

- Na altura, era uma verdadeira riqueza ter acesso a tanto texto ...
- Original no sentido de que se podia fazer o que se quisesse com o texto. Não as limitações "só para fins académicos"
- Distribuído em CD! As pessoas não confiavam na internete. As páginas desapareciam... ou não eram estáveis.

O CETEMPúblico

Agora está obviamente desatualizado. Mas continua a ser uma fonte importante para muitos trabalhos, exatamente pela sua estabilidade. (O CETENFolha ainda mais.) São uma espécie de BNC para o português.

- É importante chamar a atenção para o facto de que o texto está densamente anotado
- Por isso, podem fazer-se perguntas a muitos níveis
- Mas 180 milhões de palavras são demais para estar humanamente revisto

Exemplos de usos do CETEMPúblico

- A palavra *vela* é mais usada como substantivo ou como verbo?
- A palavra *remo* é mais frequente do que *vela*?
- Em que contextos é usada a palavra *Bragança*?
- É frequente a menção de *Chaves* e de *Bragança* na mesma frase? (e de Lisboa e Porto? E em diferentes corpos?)
- Em que contextos se usa *dramatizar*? E *desdramatizar*?
- Qual a emoção mais associada a falar?
- Qual é o verbo mais usado com *árvores*? E qual o verbo que liga mais às árvores? E o que é que as árvores mais fazem?

Árvores no CETEMPúblico

Como objeto

plantar	134
haver	74
cortar	67
derrubar	56
ter	41
ver	33
abater	22
arrancar	19
deitar	18
confundir	16
podar	14

Como sujeito

ser	67
estar	34
cair	18
plantar	16
arrancar	16
cortar	15
derrubar	13
ter	13
crescer	10
morrer	6
apresentar	6

Perguntas sobre campos semânticos

- Que mais se faz a roupa?
- Quais são as coisas mais coloridas?

Todos os usos da cor

bandeira	1276
camisola	1253
cor	1052
zona	934
cabelo	837
vinho	780
olho	712
comunidade	411
luz	398
camisa	390

Só cor cor

camisola	1239
bandeira	1237
cor	1040
zona	933
camisa	389
mancha	372
fato	366
tom	366
fundo	366
luz	352

E noutros corpos

No CONDIV (1,1 milhões de palavras sobre moda)

lã	664
seda	164
algodão	148
cor	140
veludo	113
botão	112
flanela	108
linho	108
crepe	105

No Vercial (14,8 milhões de palavras de literatura)

seda	208
céu	146
face	145
cor	124
veludo	115
roupa	114
doutor	105
lenço	92
véu	87

Comparar duas procuras no AC/DC: o Comparador

Procurar:	[pos="N.*"] [grupo="Verde"]	**Procurar:**	[pos="N.*"][grupo="Azul"]
Corpo:	CHAVE	**Corpo:**	CHAVE
Distribuir por:	lema	**Distribuir por:**	lema

☑ Mostrar totais ☐ Fundir numa única tabela

Limite mínimo de frequência 10

procurar

Total	2975
zona	237
cor	65
'paço	62
bandeira	44
mancha	38
linha	34
folha	33
camisola	24

Total	1677
bandeira	174
céu	82
cartão	69
cor	48
fundo	39
camisa	37
luz	32
fato	24

Nuvens de palavras no corpo JMG

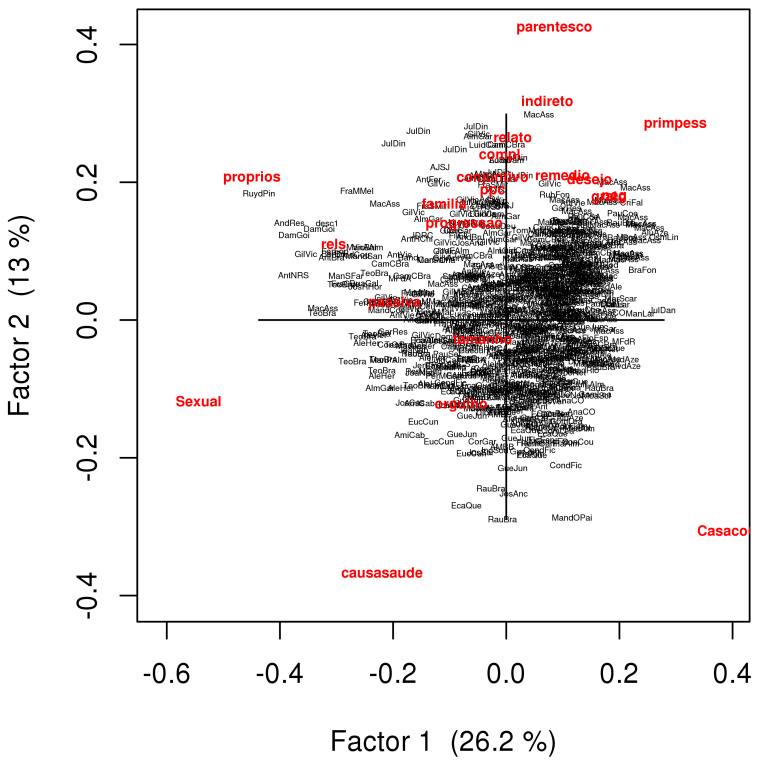
Lemas do sentimento infelicidade

Lemas do sentimento admiração

Serviços da Linguateca

Vários tipos de serviços

- Ensinador
- WebJspell
- Folheador
- Corpógrafo
- Literateca



Entremezzo: Quantas pessoas viram este filme?

https://www.youtube.com/watch?v=CtbfmSPB_Fk

Folheador

FOLHEADOR

Palavra ou Termo 1: convento Termo 2: Relação a procurar:

> Todas < Pesquisar

A procurar pela palavra: **"convento"**.

*Apresentados 10 resultados de **45** no total.*

	TRIPLOS			RECURSO(S)	GRAU DE CONFIANÇA	
	TERMO1	RELAÇÃO	TERMO2	TODOS / OK	SIMPLES	COMPOSTA
	convento *(nome)*	FINALIDADE_DE	dote *(nome)*	wiki, da, papel	6	0.0
	convento *(nome)*	SINONIMO_N_DE	claustro *(nome)*	wiki, da, papel	71	0.0
	convento *(nome)*	SINONIMO_N_DE	mosteiro *(nome)*	ot, tep, da	84	0.0
	convento *(nome)*	SINONIMO_N_DE	mosteiro *(nome)*	ot, tep, papel	84	0.0
	convento *(nome)*	SINONIMO_N_DE	abadia *(nome)*	ot, tep	4	0.0
	convento *(nome)*	HIPERONIMO_DE	freiria *(nome)*	da, papel	0	0.0
	convento *(nome)*	HIPONIMO_DE	casa *(nome)*	wiki, papel	134	0.0
	convento *(nome)*	SINONIMO_N_DE	reclusão *(nome)*	wiki, papel	8	0.0
	convento *(nome)*	SINONIMO_N_DE	casarão *(nome)*	wiki, papel	1	0.0
	convento *(nome)*	SINONIMO_N_DE	abadia *(nome)*	ot, tep	4	0.0

2 3 4 ... fim >

Última atualização: 18 de Maio de 2012

Perguntas, comentários e sugestões

Corpógrafo

Um sistema que foi gizado pela Belinda Maia com base nas necessidades concretas dos seus alunos e das suas aulas, centrado no utilizador.

- obter facilmente pequenos corpos "para usar e deitar fora"
- transformar os textos num formato processável
- permitir a gestão dos dados de uma forma privada
- permitir corpos comparáveis para estudos de tradução e terminologia

Um ambiente que é usado em muitos locais do mundo para aprender terminologia, e que foi centrado nas necessidades reais de um grupo grande de pessoas.

Ensinador

Um serviço para aqueles que ensinam português, para criarem facilmente exercícios (lexicais ou gramaticais) baseados nos corpos do AC/DC.

- Permite uma escolha criteriosa dos casos
- Dá automaticamente (uma) solução
- Tem um conjunto relativamente simples de extensões em relação ao próprio AC/DC

Concordância temporal: *quando* ou *enquanto*?

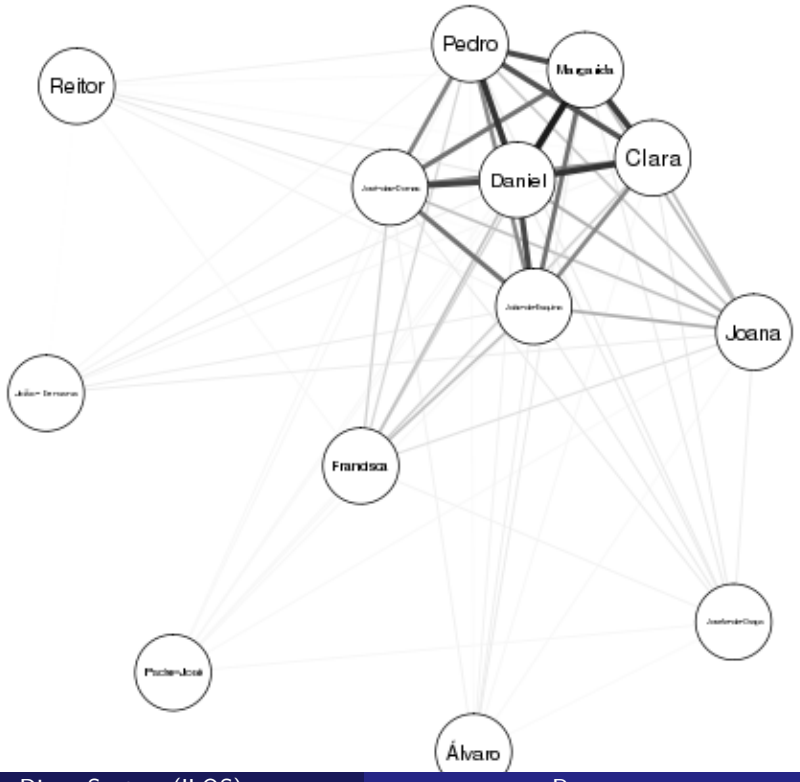
1. *FSP940101-096*: Um som indica que a bola caiu fora, _____ a informação é enviada ao árbitro de cadeira por computador .
2. *FSP940102-111*: Gilvane Ribeiro, uma cearense de 20 anos, com quem Chemwoyo namorou _____ esteve em São Paulo em 92, reapareceu anteontem .
3. *FSP940104-112*: Francisco Maturana trabalhava como dentista do Nacional de Medellin, em 1986, _____ ele e o então mister do clube, o ex-craque uruguaio Luís Cubilla, se tornaram amigos .

WebJspell

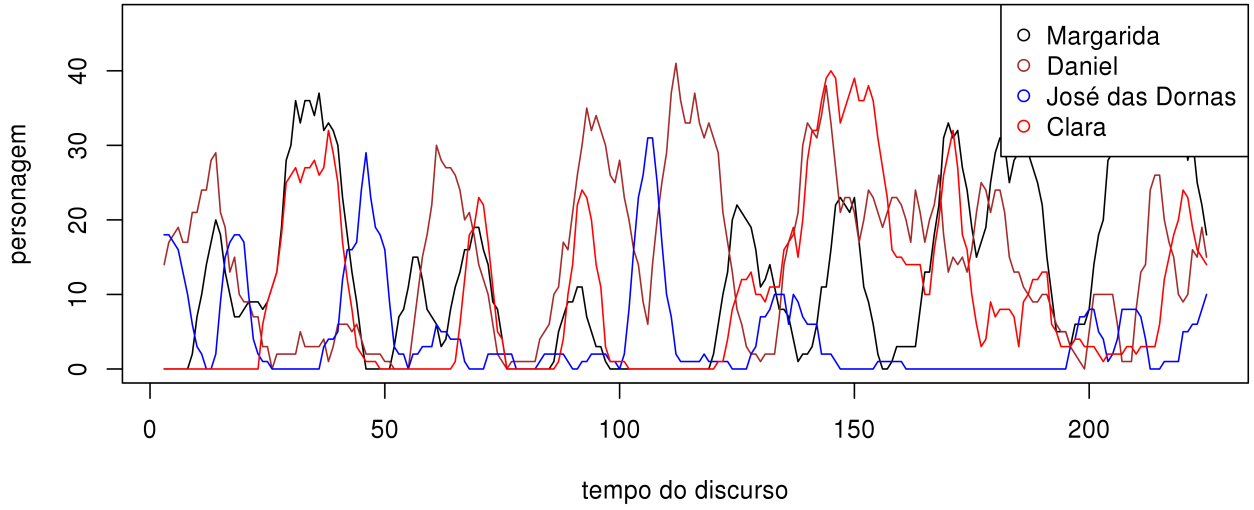
Criado pelo pólo de Braga da Linguateca, um serviço que permite fazer a correção ortográfica de páginas da internete em português, ou simplesmente fazer análise morfológica com base em vários dicionários.

Literateca

O último grito, tentando analisar obras literárias a distância:



Literateca



Conclusão

- A Linguateca não teve nunca como objetivo criar recursos ou serviços para o grande público, os falantes de língua portuguesa. Ou seja, não teve como destinatários a presente audiência, nem como objetivo *sensibilizar para o uso rigoroso da língua*.
- Contudo, os recursos e serviços criados podem potenciar um maior conhecimento da língua e permitir um desenvolvimento individual no contacto com esta
- Poderão também permitir questões e surpresas que agucem a curiosidade sobre o nosso meio comum de expressão, e levem a trabalhos de investigação

Referências

- Costa, Albert, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner & Boaz Keysar. “Your Morals Depend on Language". *PLOS ONE*, April 2014, Volume 9, Issue 4.
- Santos, Diana & Belinda Maia. “Language, emotion, and the emotions: A computational introduction", *Language and Linguistics compass*, 12, 5, 2018.
- Santos, Diana. “Português internacional: alguns argumentos". In José Teixeira (ed.), *O Português como Língua num Mundo Global: problemas e potencialidades*, Universidade do Minho, 2016, pp. 49-66.
- Santos, Diana. “Para documentar o "ministro da língua"”. In Catalão Alves (ed.), *Caminhos do Conhecimento, O Legado de José Mariano Gago. Dia Nacional dos Cientistas*, Ciência Viva, 2018, pp. 147-161.
- Santos, Diana. “Tradução técnica, terminologia e criatividade”. In Aparecida Negri Isquerdo & Giselle Olivia Mantovano Dal Corno (eds.), *As Ciências do léxico: Lexicologia, Lexicografia, Terminologia*, Vol. VIII, Editora da UFMS, Campo Grande, MS, 2018, pp. 253-272.

Artigos que descrevem alguns dos recursos

- Sarmento, Luís, Belinda Maia, Diana Santos, Luís Cabral & Ana Sofia Pinto. "Corpógrafo V3: From simple word-concordance to semi-automatic knowledge engineering". In Nicoletta Calzolari et al. (eds.), *Proceedings of LREC 2006 (LREC'2006)* (Génova, 22-28 de Maio de 2006), pp. 1502-5.
- Gonçalo Oliveira, Hugo, Hernâni Costa & Diana Santos. "Folheador: browsing through Portuguese semantic relations", *Proceedings of EACL'2012* (Avignon, France, April 23-27, 2012), pp. 35-40.
- Inácio, Susana, Diana Santos & Rosário Silva. "COMPARAndo cores em português e inglês". In Sónia Frota & Ana Lúcia Santos (orgs.), *Artigos seleccionados do XXIII Encontro da Associação Portuguesa de Linguística* (Évora, 1-3 de Outubro de 2007), APL, 2008, pp. 271-86.

Artigos que descrevem alguns dos recursos

- 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.
- Simões, Alberto & Diana Santos. "Nos bastidores da Gramateca: uma série de serviços", *Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish, at PROPOR 2014, São Carlos, Brazil*, 9 de outubro de 2014, pp. 97-104.
- Santos, Diana & Cristina Mota. "A admiração à luz dos corpos". In Simões et al. (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam. Homenagem a Belinda Maia*, OSLa, Vol 7, No 1 (2015), pp. 57-77.

O maior recurso da Linguateca são as pessoas!

- Muitas delas continuam o trabalho em áreas que começaram na Linguateca
- Outras dedicam-se a novas áreas
- Outras ainda continuam a fazer parte da Linguateca desenvolvendo projetos e recursos sob essa bandeira comum
- Novas pessoas entram em contacto e querem colaborar ou desenvolver novos recursos

Quem nos faz continuar, são todos os utilizadores dos nossos serviços e recursos, e os entusiastas da língua portuguesa, no mundo digital, e não só!

Origem das imagens

http://elarboldorado.com/wp-content/uploads/2015/01/imaginar-cosas-positivas.jpg
https://commons.wikimedia.org/wiki/File:Janela_g%C3%B3tica_Bragan%C3%A7a.jpg
https://pt.wikipedia.org/wiki/L%C3%ADngua_portuguesa#/media/Ficheiro:Lusophone_World.svg