Pluralidades na cor: contrastando a língua do Brasil e de Portugal

Diana Santos
Rosário Silva
Cláudia Freitas
Augusto Soares da Silva

The world of colour in Portuguese
do Brazil and Portugal differ?
A corpus-based study

Diana Santos
Rosário Silva
Cláudia Freitas
Augusto Soares da Silva

Goals of this presentation

- Present AC/DC for variety studies and semantic studies of Portuguese
- Discuss some data
- Discuss methodologies and conclusions
- Raise interest on corpus-based semantic field comparison

It all began…

- In 1998 with a preparatory project to improve the computational processing of Portuguese that led to *Linguateca* (2000-)
  - One of the oldest and most used achievements is the AC/DC project
- In 2005-6 when CONDIVport was included in AC/DC

- In 1990 when I started work in semantics…

Context

- The AC/DC corpora, including CONDIVport
- The corte-e-costura program for human-revised semantic annotation
- Previous work on colour under the scope of COMPARA

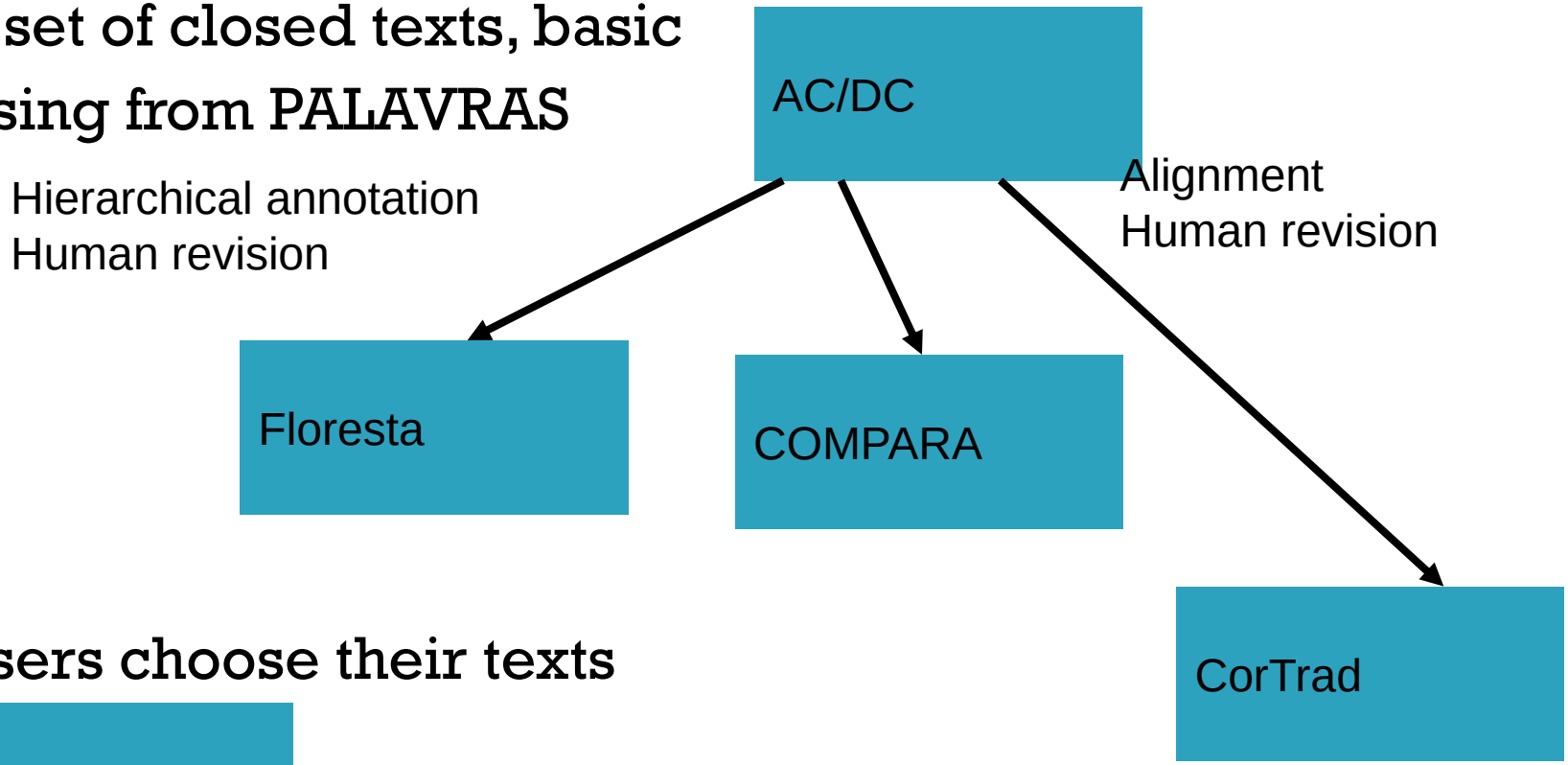
- Synonym aid and lexical ontologies for Portuguese

A bird's eye view on AC/DC

- Slides from last year's presentation

Similarities and differences in Linguateca corpora

- A set of closed texts, basic parsing from PALAVRAS



- Users choose their texts

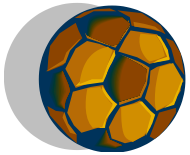
Corpógrafo

Corpus gallery in the AC/DC cluster

- General newspapers
  - CETEMPúblico
  - CETENFolha (→ São Carlos)
  - CHAVE
  - Notícias de Moçambique

- Regional newspapers
  - NatMinho
  - DiaCLAV
  - Diário Gaúcho

- Specific newspapers
  - Sports : CONDIVport
  - Political: Avante!
  - Fashion: CONDIVport
  - Health: CONDIVport
  - Science: CorTradjorn



- Literary
  - Vercial
  - ClassLPPE
  - ENPCpub
  - COMPARA
  - CorTradlit

Adapted from Rocha (2007)

Corpus gallery in the AC/DC cluster (cont.)

- Oral documents
  - Museu da Pessoa
  - ECI-EBR falado
  - Selva falado

- Technical
  - CorTradtec
  - ECI-EE
  - NILC/São Carlos tec
  - Selva Ciência

- Evaluation resources
  - CDHAREM
  - AmostRA
  - FrasesPP

- Email
  - Listas: ANCIB
  - SPAM: CoNE

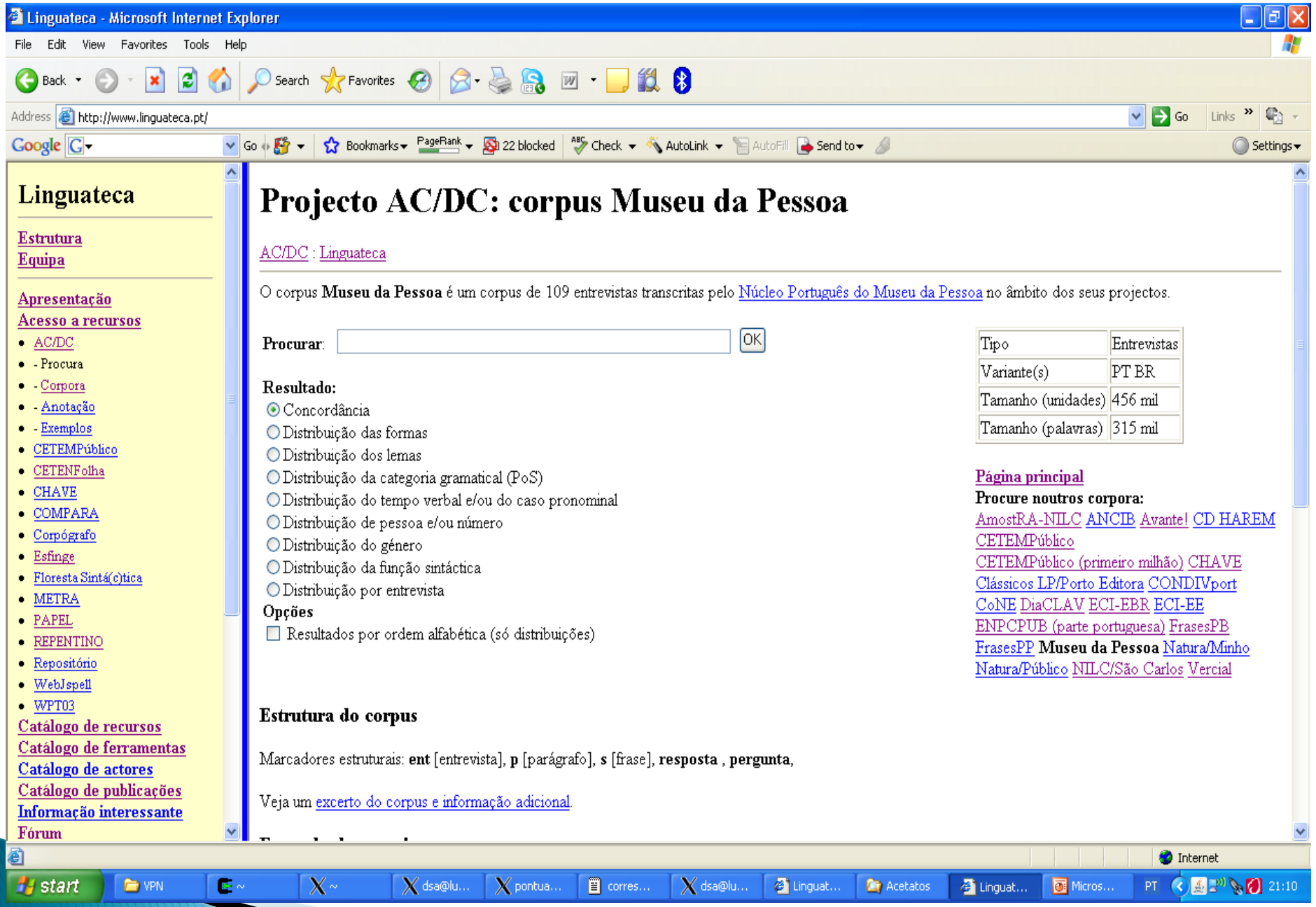
- Web
  - Amazônia

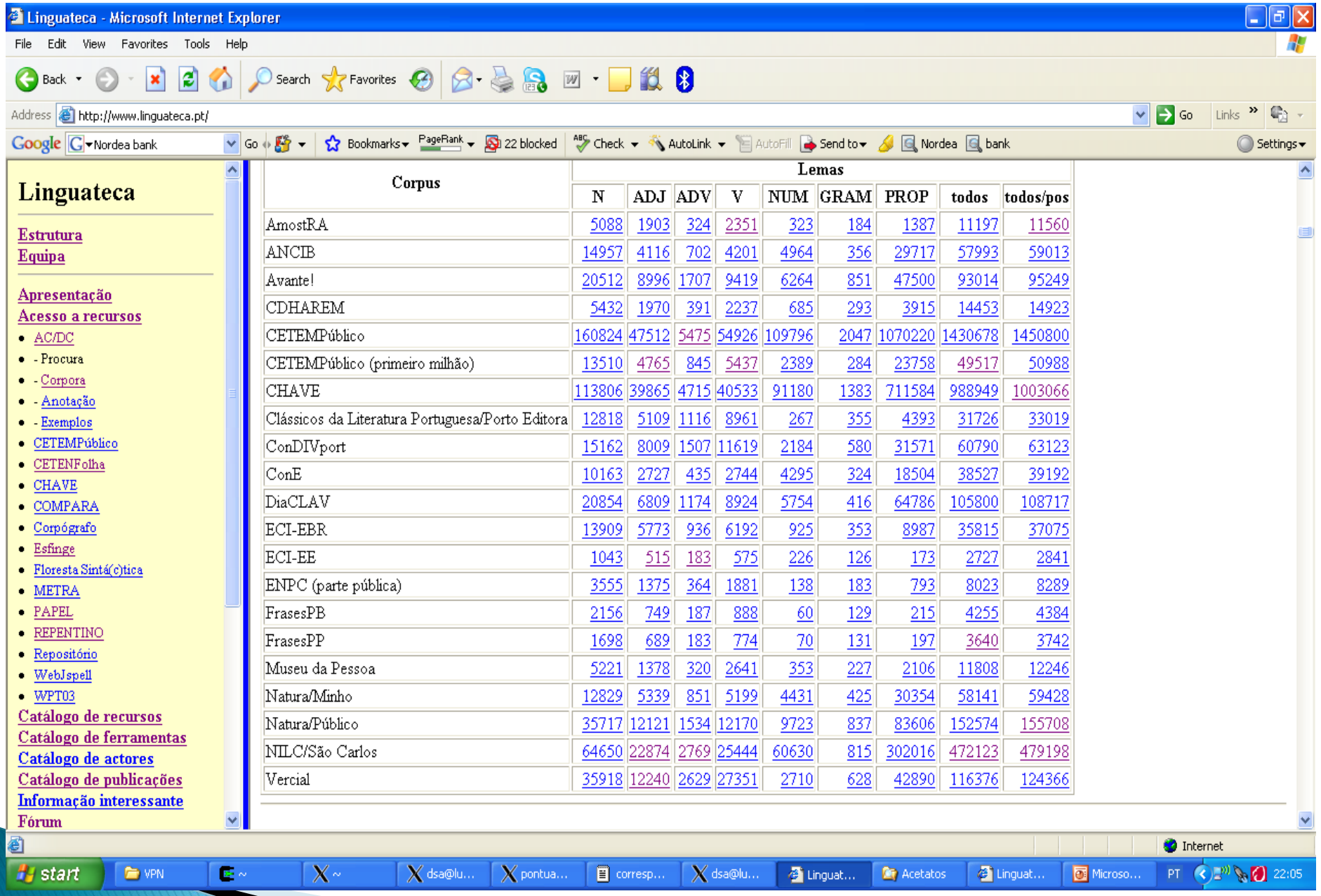
- "Historical"
  - CETEMPúblico (primeiro milhão)
  - NatPublico

Adapted from Rocha (2007)

Brief description of AC/DC

- **A**cesso a **C**orpora / **D**isponibilização de **C**orpora
- Ca. 20 different corpora
- Ca. 360 million words, 16 million sentences
- Portuguese and Brazilian varieties, a few other texts from others
- Different genres, mainly contemporary
- Perl interface to the IMS (Open) CWB (corpus workbench)
- Common tokenization
- Use of the PALAVRAS parser (Bick, 2000) for linguistic annotation
- (Semi-automatic) annotation of selected semantic features

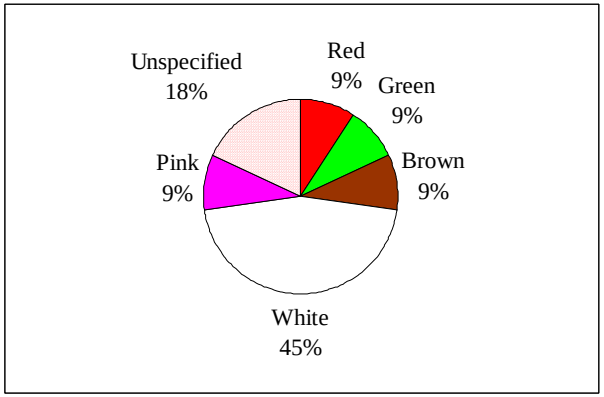

Linguateca - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.linguateca.pt/
Go Links
Google Go Bookmarks PageRank 22 blocked Check AutoLink AutoFill Send to Settings
Linguateca
Estrutura
Equipa
Apresentação
Acesso a recursos
AC/DC
- Procura
- Corpora
- Anotação
- Exemplos
CETEMPúblico
CETENFolha
CHAVE
COMPARA
Corpógrafo
Esfinge
Floresta Sintá(c)tica
METRA
PAPEL
REPENTINO
Repositório
WebJspell
WPT03
Catálogo de recursos
Catálogo de ferramentas
Catálogo de actores
Catálogo de publicações
Informação interessante
Fórum
Projecto AC/DC: corpus Museu da Pessoa
AC/DC : Linguateca
O corpus Museu da Pessoa é um corpus de 109 entrevistas transcritas pelo Núcleo Português do Museu da Pessoa no âmbito dos seus projectos.
Procurar: OK
Resultado:
Concordância
Distribuição das formas
Distribuição dos lemas
Distribuição da categoria gramatical (PoS)
Distribuição do tempo verbal e/ou do caso pronominal
Distribuição de pessoa e/ou número
Distribuição do género
Distribuição da função sintáctica
Distribuição por entrevista
Opções
Resultados por ordem alfabética (só distribuições)
Tipo Entrevistas
Variante(s) PT BR
Tamanho (unidades) 456 mil
Tamanho (palavras) 315 mil
Página principal
Procure noutros corpora:
AmostRA-NILC ANCIB Avante! CD HAREM
CETEMPúblico
CETEMPúblico (primeiro milhão) CHAVE
Clássicos LP/Porto Editora CONDIVport
CoNE DiaCLAV ECI-EBR ECI-EE
ENPCPUB (parte portuguesa) FrasesPB
FrasesPP Museu da Pessoa Natura/Minho
Natura/Público NILC/São Carlos Vercial
Estrutura do corpus
Marcadores estruturais: ent [entrevista], p [parágrafo], s [frase], resposta , pergunta,
Veja um excerto do corpus e informação adicional.
Internet
start VPN dsa@lu... pontua... corres... dsa@lu... Linguat... Acetatos Linguat... Micros... PT 21:10

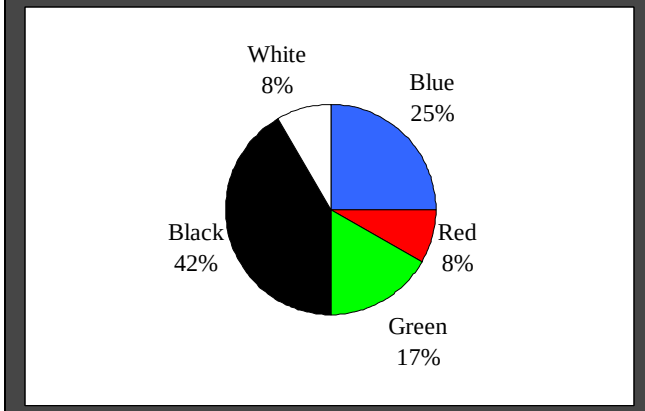


Corpus	Lemas								
	N	ADJ	ADV	V	NUM	GRAM	PROP	todos	todos/pos
AmostRA	5088	1903	324	2351	323	184	1387	11197	11560
ANCIB	14957	4116	702	4201	4964	356	29717	57993	59013
Avante!	20512	8996	1707	9419	6264	851	47500	93014	95249
CDHAREM	5432	1970	391	2237	685	293	3915	14453	14923
CETEMPúblico	160824	47512	5475	54926	109796	2047	1070220	1430678	1450800
CETEMPúblico (primeiro milhão)	13510	4765	845	5437	2389	284	23758	49517	50988
CHAVE	113806	39865	4715	40533	91180	1383	711584	988949	1003066
Clássicos da Literatura Portuguesa/Porto Editora	12818	5109	1116	8961	267	355	4393	31726	33019
ConDIVport	15162	8009	1507	11619	2184	580	31571	60790	63123
ConE	10163	2727	435	2744	4295	324	18504	38527	39192
DiaCLAV	20854	6809	1174	8924	5754	416	64786	105800	108717
ECI-EBR	13909	5773	936	6192	925	353	8987	35815	37075
ECI-EE	1043	515	183	575	226	126	173	2727	2841
ENPC (parte pública)	3555	1375	364	1881	138	183	793	8023	8289
FrasesPB	2156	749	187	888	60	129	215	4255	4384
FrasesPP	1698	689	183	774	70	131	197	3640	3742
Museu da Pessoa	5221	1378	320	2641	353	227	2106	11808	12246
Natura/Minho	12829	5339	851	5199	4431	425	30354	58141	59428
Natura/Público	35717	12121	1534	12170	9723	837	83606	152574	155708
NILC/São Carlos	64650	22874	2769	25444	60630	815	302016	472123	479198
Vercial	35918	12240	2629	27351	2710	628	42890	116376	124366

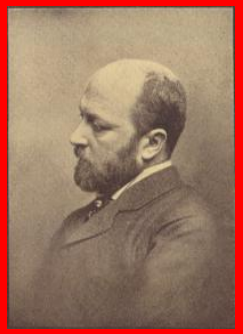
COMPARA: (EN) Author with highest % of colour:

Lewis Carrol

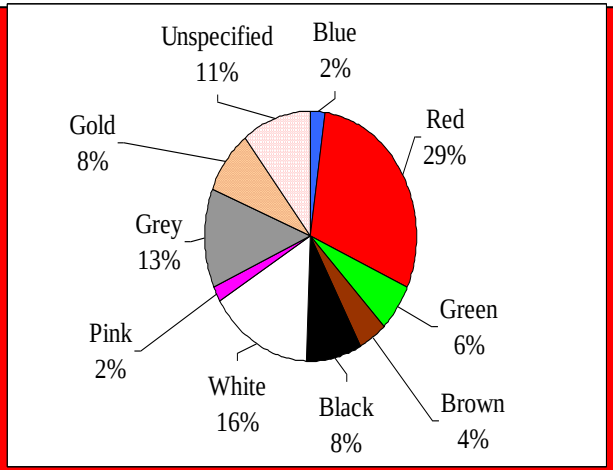


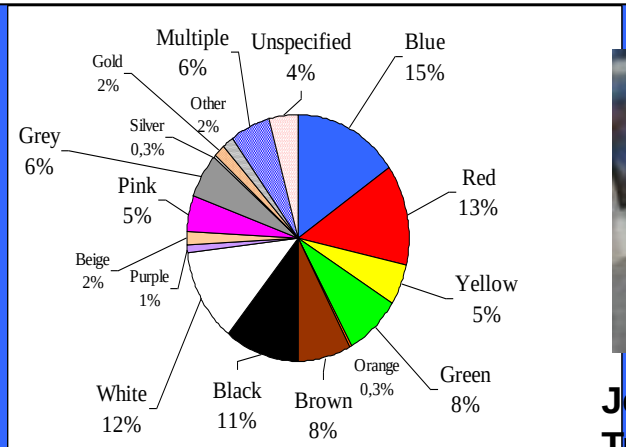


Mary Shelley



Henry James



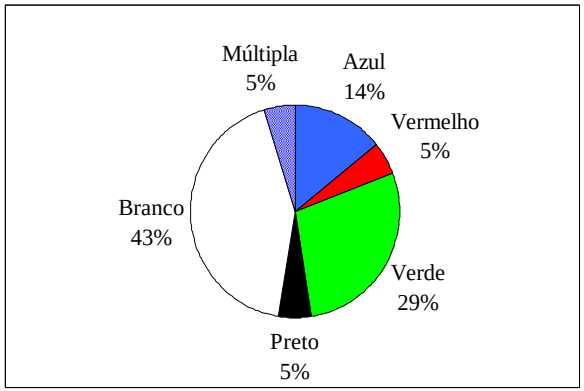


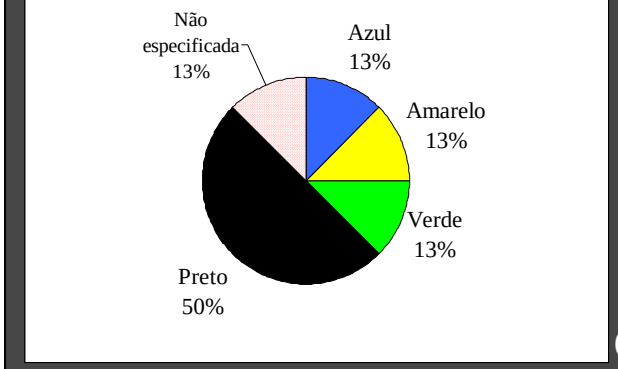
Joanna Trollope

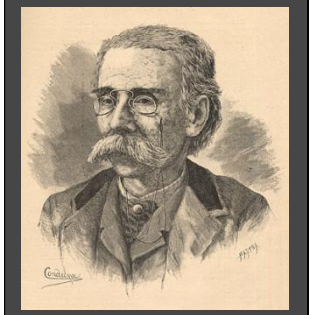
Silva, Inácio & Santos (2008)

COMPARA: (PT) author with highest % of colour:

José de Alencar



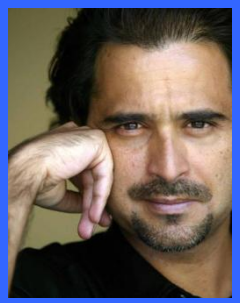




Camilo Castelo

Mia Couto

26%



José Eduardo Agualusa

31%

Jorge de Sena

24%

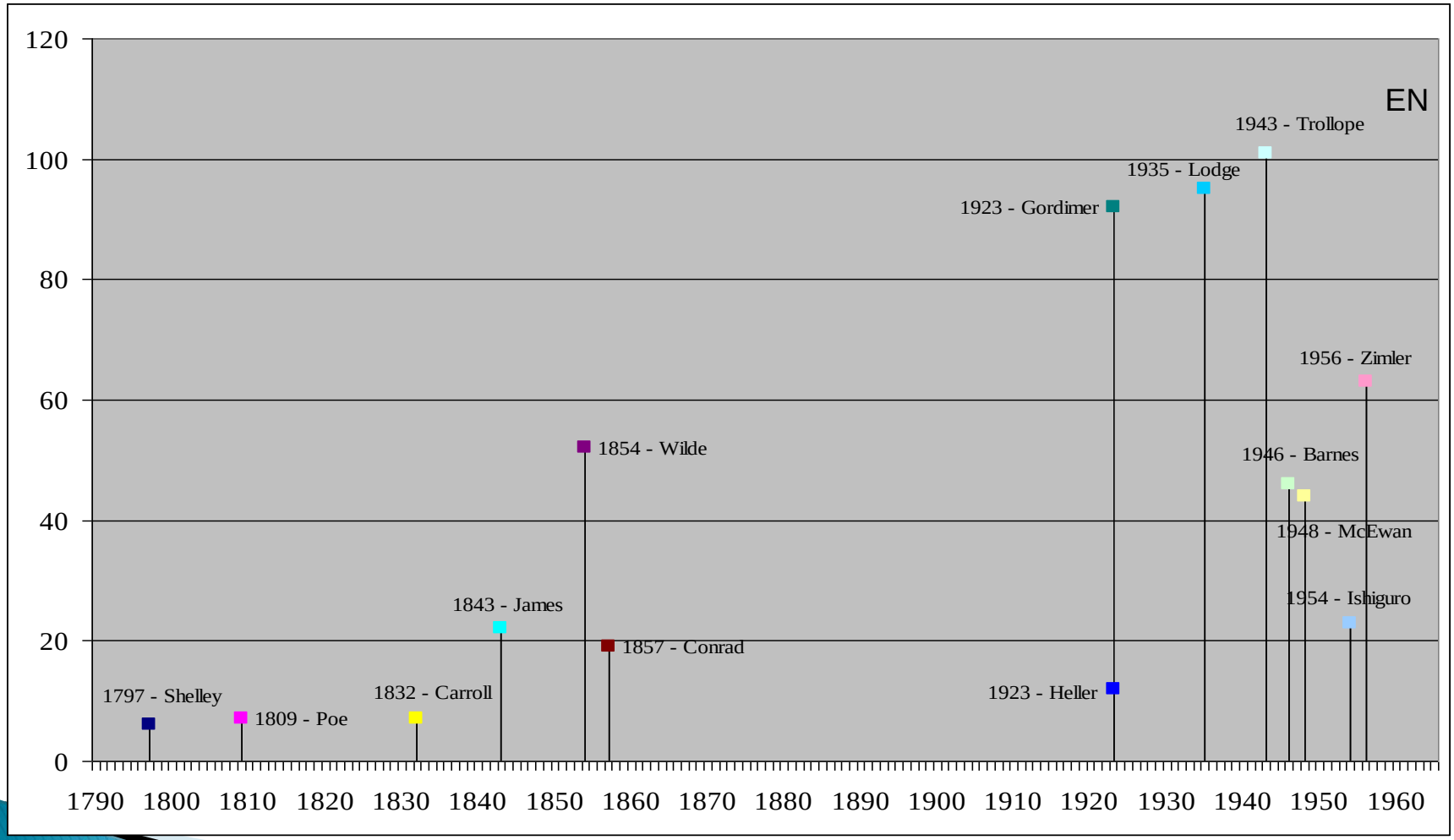
Marcos Rey

44%

Silva, Inácio & Santos (2008)

COMPARA: Does colour quantity change with time?

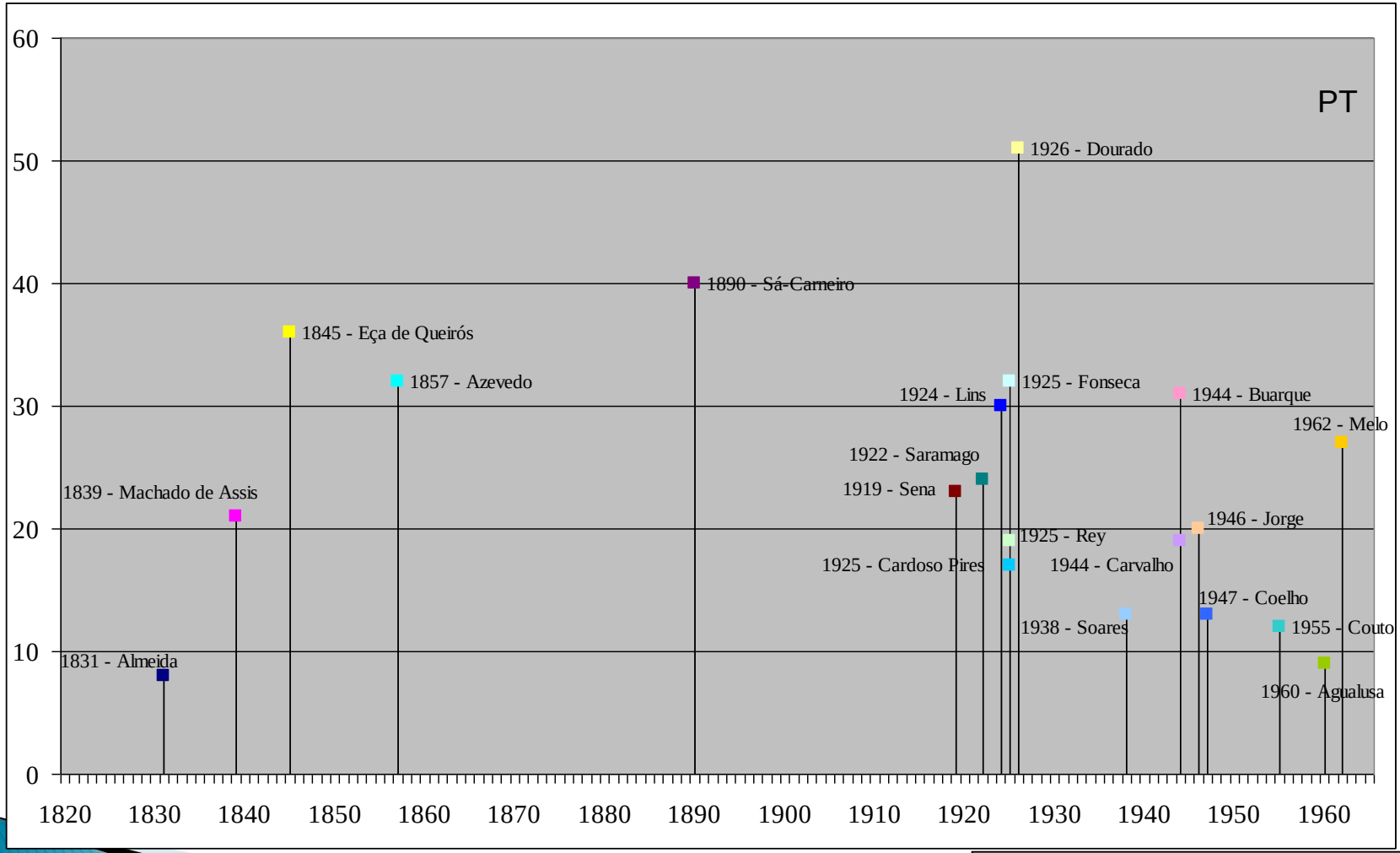
Number of colour types per authors' birth date (English-speaking authors)



Silva, Inácio & Santos (2008)

COMPARA: Does colour quantity change with time?

Number of colour types per authors' birth date (Portuguese-speaking authors)



Silva, Inácio & Santos (2008)

What does "semantic annotation" mean?

- In the AC/DC context
- Choose a number of semantic tags for particular domains, and annotate all text with them
- Batch/Interactive process, with a human on the loop, with the goal of having 100% correct annotation
  - Lexical information
  - Rule application
- All choices taken in the annotation documented

The semantic annotation process

- "General" rules
  - Appropriate for many contexts, general enough to be applied to many themes, subjects and genres
  - Rule-like flavour
- Corpus-specific rules
  - Cases which are like this contingently
- To end up with a 100% accurate annotation
- Not necessarily easy to decide where to place a particular rule
  - promotions and depromotions occur frequently

The colour lexicons

- Colour in general: *azul*, *amarelo*
- Colour just in some POS uses
  - Adjective: *laranja*, *castanho*
- Colour in quite rare situations
  - Because of ambiguity: *alvo*, *louro*, *creme*
  - Because the word has not (yet?) lost its main/original meaning: *café*, *cinza*
- Inherently vague colour words: *ouro*, *verde*
- Colour words in specific areas: *moreno*, *tinto, marronzinho*
- Metaphorical: *branqueamento*, *cinzentão*

A multiword part

- Colours with more than one word: *cor de rosa*, *peito de rola*, *verde claro*
- Cor:original
  - Fixed expressions whose main point is not colour: *páginas amarelas*, *zonas verdes*, *cartões amarelos*, *papel pardo*
  - Metaphorical use: *vida negra*, *sorriso amarelo*
  - Metonymical use: *capacetes azuis*, *governo laranja*
  - Specialized uses in specific domains: *carnes brancas*, *feijão verde, sabão azul*

Colour properties with more or less than one colour

- *de cor*, *colorida* ,*tricolor* cor:Nãoespecificada
- *incolor*, *transparente*, *sem cor*, *de cor indefinida* cor:Ausência
- *bandeira azul e branca* cor:Múltipla
- *equipa verde-rubra* cor:Múltipla

Grouping

- True colours (denoting mainly visual properties)
  - 19 groups (14 of a single colour: BRANCO, PRETO, AZUL, AMARELO, VERMELHO, LARANJA, VERDE, ROXO, CASTANHO, CREME, CINZENTO, ROSA, DOURADO, PRATEADO)
  - Different groups: Outras, Ausência, Desconhecido, Múltipla, Nãoespecificado
- "Colours" associated with
  - human race (branco, preto, negro, amarelo, …)
  - human appearance (loiro, moreno, grisalho, …)
  - wine (branco, tinto, verde)
  - politics (verdes, laranjas, vermelhos, …)
  - sports teams
- "Colours" associated with maturity (lack of): *verde*
- Other uses (cor:original)

Lexicon size (1730 types)

- (Almost) always colours (N or A): 1582
  - Single words: 1118
  - Multiword expressions: 464
- Only when adjectives: 47
- Verbs: 73
- Possible colours: 29
  - Single words: 24
  - Multiword expressions: 5
- Domain related: humana (101) raça (9) vinho (1) equipa (10)
- Metaphorical -- original: 61
  - Single words: 27
  - Multiword expressions: 31

The annotation in context

- Number of colour tokens in the corpora of the AC/DC cluster-- full data in the file http://www.linguateca.pt/acesso/ArcoIris.pdf
- Avante: 2,675
- NILC/São Carlos: 28,302
- CHAVE: 82,571
- CONDIVport: 20,435
- …
- Total: 328,633

The annotation in context

- How many different types? 1070
- By lemma (only pure colour):

dourar: 2175
rosa: 2203
dourado: 2364
colorir: 3223
encarnado: 3288
colorido: 3589
cinzento: 3890
laranja: 4635
amarelo: 9368
preto: 14101
vermelho: 16657
azul: 17101
verde: 21120
cor: 30824
negro: 38208
branco: 39348

The annotation in context

- How many different types? 1092
- By lemma (all colours):

branco: 42663
negro: 41803
cor: 33993
verde: 22882
azul: 18776
vermelho: 17937
preto: 15196
amarelo: 9528
laranja: 4641
cinzento: 4097
colorido: 3780
encarnado: 3347
colorir: 3331
dourar: 3067
dourado: 2750
alvo: 2604
rosa: 2343

The annotation in context

- How many different groups? 136
- By group (pure colours):

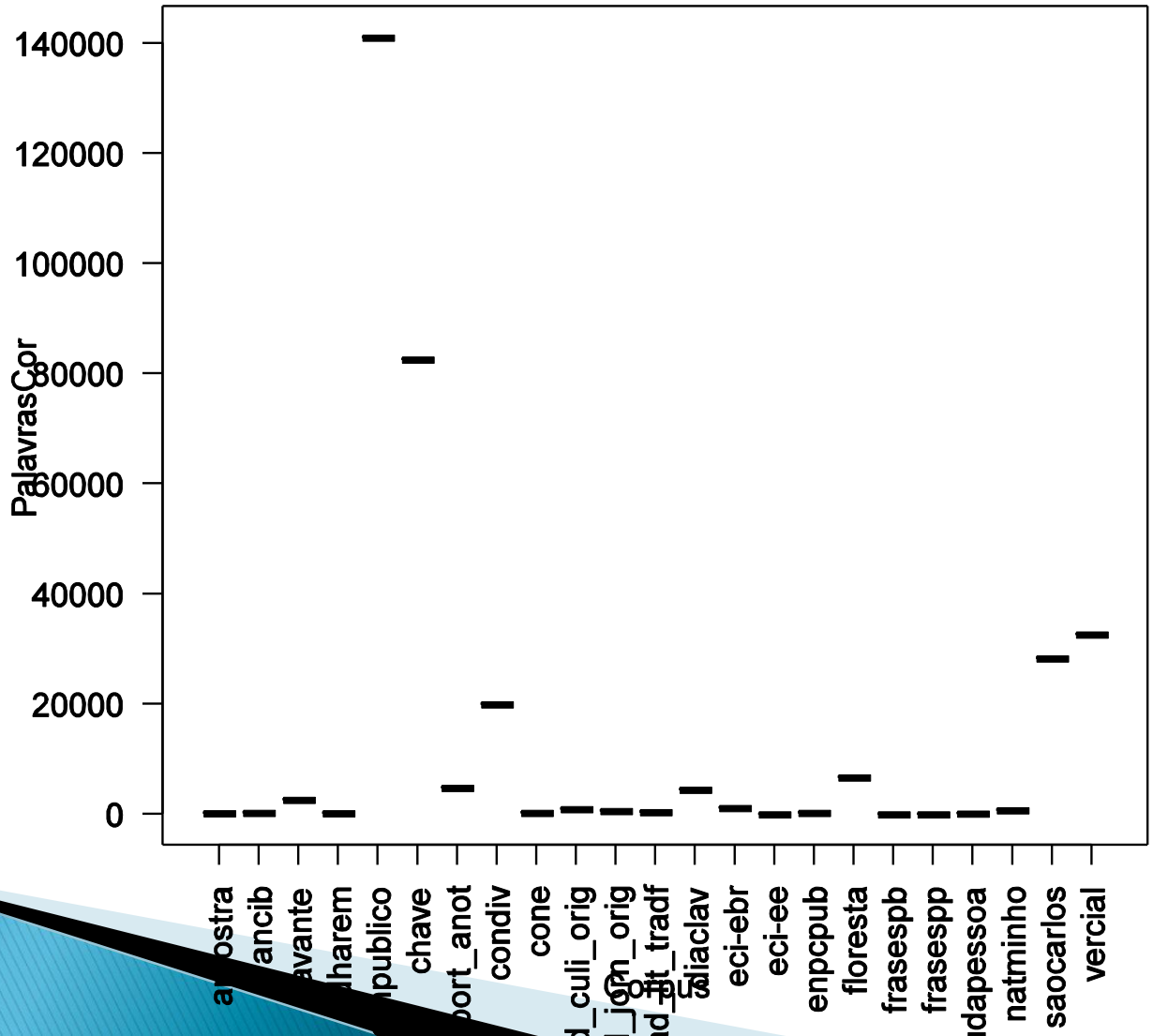
Preto: 52965
Branco: 46426
Nãoespecificada: 38505
Verde: 22936
Vermelho: 22370
Azul: 19203
Amarelo: 10279
Rosa: 6297
Cinzento: 6145
Laranja: 5230
Dourado: 4872
Castanho: 3079
Roxo: 1762
Múltipla: 1569
Creme: 1106

OutrasCores:gerâneo: 1
OutrasCores:adamascado: 1
OutrasCores:gelo: 1

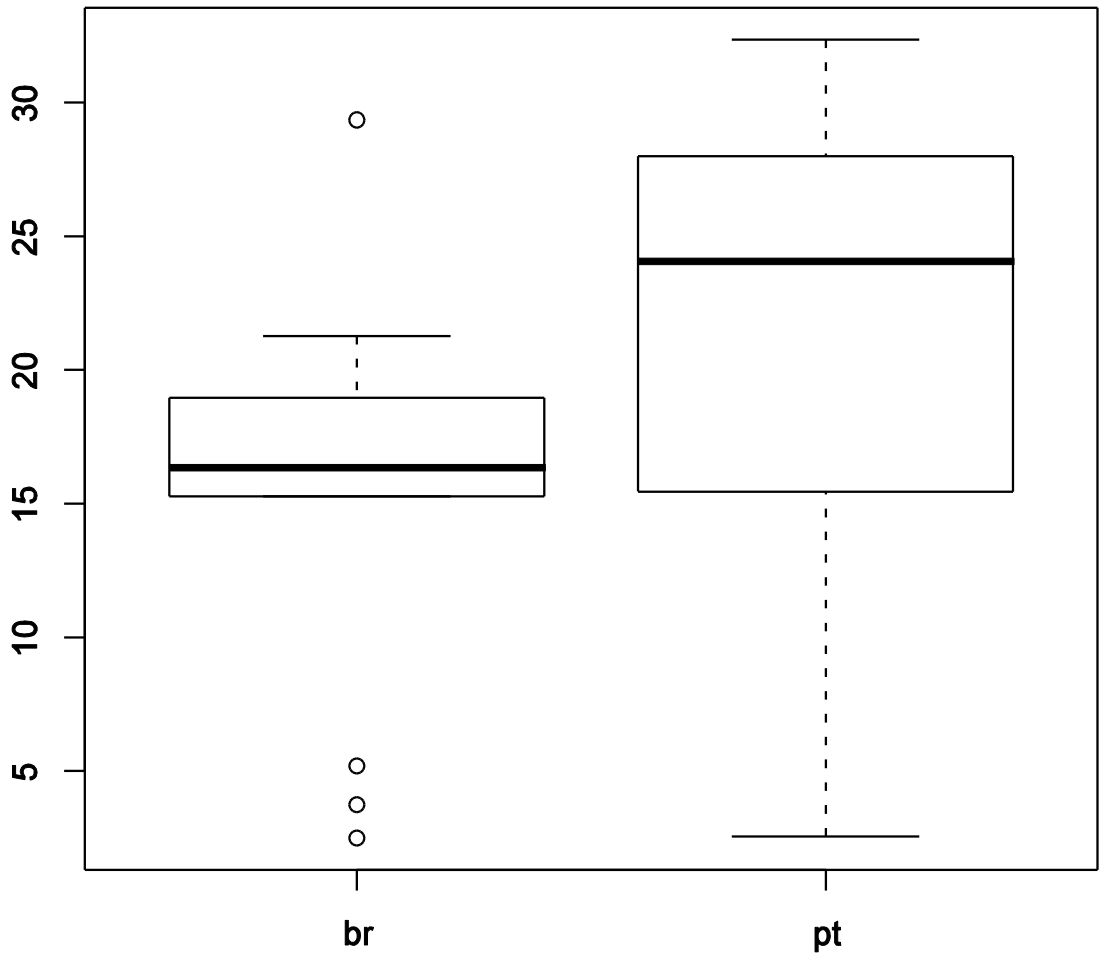
Kinds of decisions for annotation in context

- How many colours? Which lemma?
- *Rímel **colorido azul marinho** ou castanho na ponta dos cílios também dão cor*
- *A cor dominante é o azul – marinho ou **ultramarino**, conforme a sensibilidade de cada um*
- *As cores vão da gama dos verdes, aos brancos óptico e **marfim***
- Which group?
- *Nem pelo formato, nem pela **cor** do papel, nem pela impressão é o Oslobodenje que conhecem há quase cinquenta anos .*

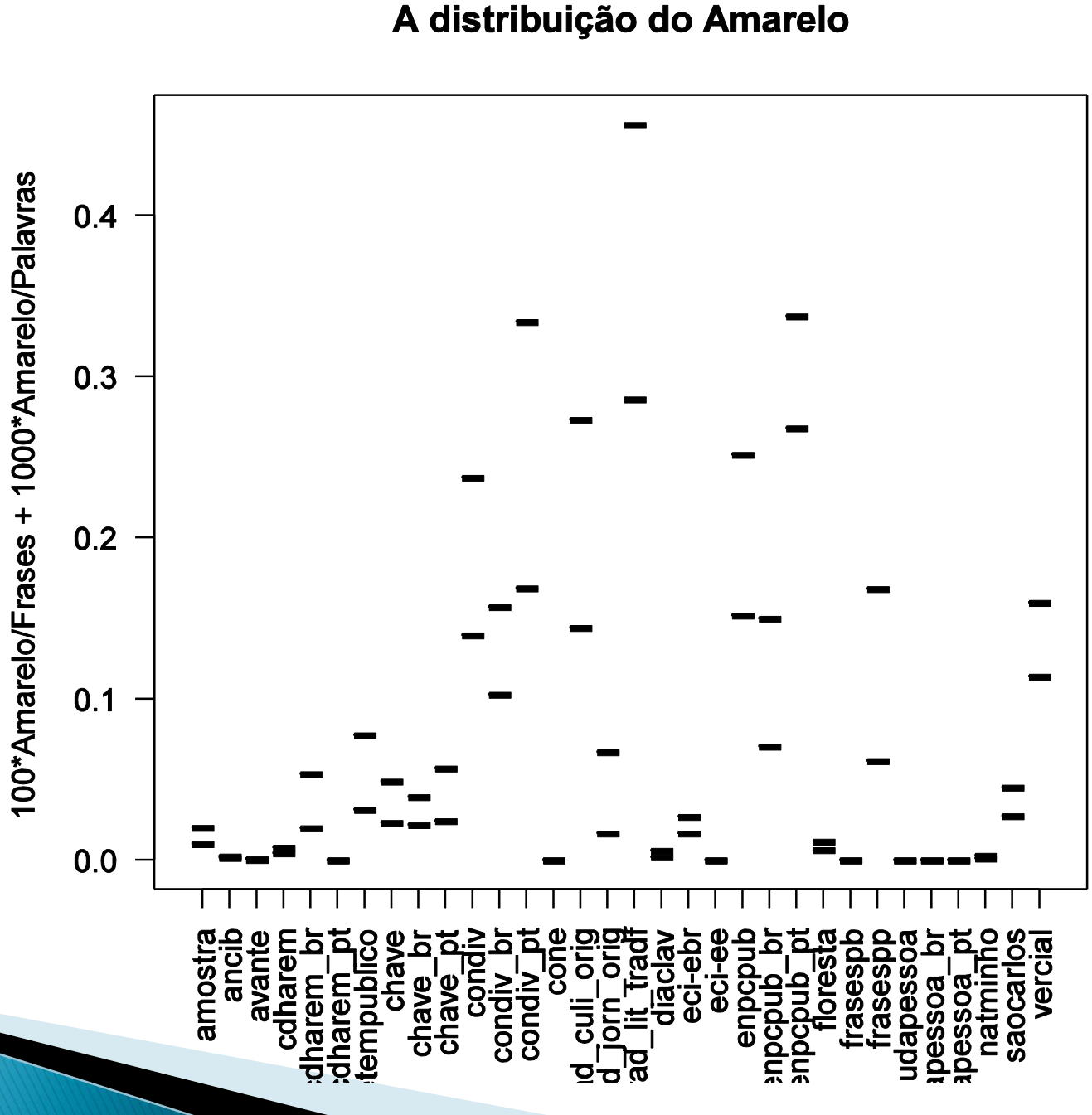
Preliminary explorations (1)



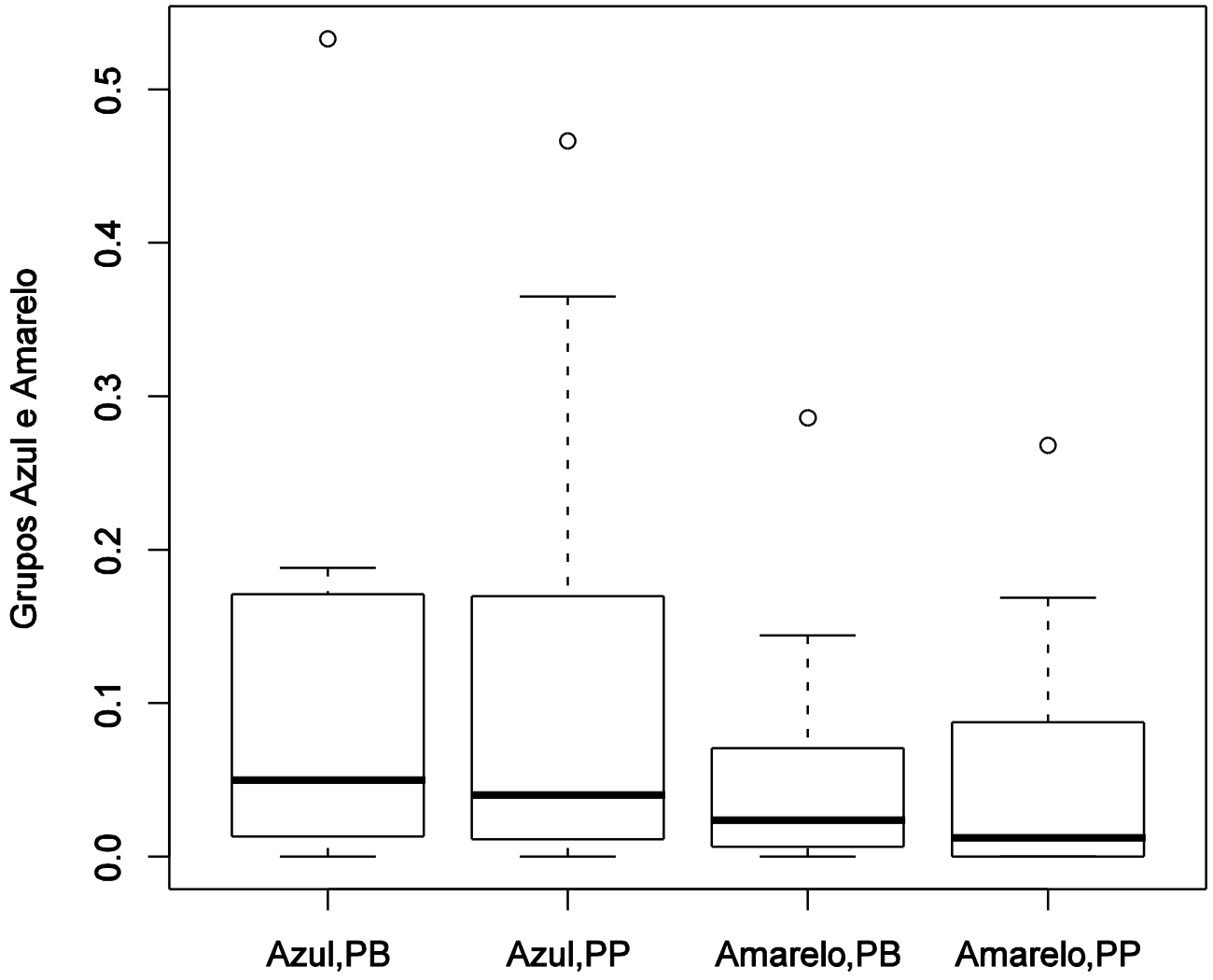
(2)



(3)


A distribuição do Amarelo
100*Amarelo/Frases + 1000*Amarelo/Palavras
0.0
0.1
0.2
0.3
0.4
amostra
ancib
avante
cdharem
cdharem_br
cdharem_pt
chave
chave_br
chave_pt
condiv
condiv_br
condiv_pt
cone
eci-ebr
eci-ee
enpcpub
floresta
frasespb
frasespp
natminho
saocarlos
vercial

(4)


Grupos Azul e Amarelo
0.0
0.1
0.2
0.3
0.4
0.5
Azul,PB
Azul,PP
Amarelo,PB
Amarelo,PP

Comparing groups in two varieties

		PRETO	BRANCO	VERMELHO	AZUL	VERDE	AMARELO	LARANJA
PT	CONDIV	1318	2336	1037	1209	733	590	150
	CHAVE	8448	6254	3192	3711	3773	1461	589
Total PT		**9766**	**8590**	**4229**	**4920**	**4506**	**2051**	**739**
BR	CONDIV	765	829	695	423	299	254	39
	CHAVE	6504	4308	1796	1247	1639	859	175
Total BR		**7269**	**5137**	**2491**	**1670**	**1938**	**1113**	**214**
TOTAL		**17035**	**13727**	**6720**	**6590**	**6444**	**3164**	**953**

CONDIV: PT 3,284,575 (55.5%) BR 2,631,558 (44.5%)
CHAVE: PT 54,947,072 (60.5%) BR 35,699,765 (40%)

Biderman et al (2007)

- Biderman, Maria Tereza Camargo, Maria Fernanda Bacelar do Nascimento & Luisa Alice Santos Pereira. "Uso das cores no português brasileiro e no português europeu". In Aparecida Negri Isquierdo & Ieda Maria Alves (eds.), *As ciências do léxico: Lexicologia, lexicografia, terminologia*, vol. III, Editora UFMS, Associação editorial humanitas, 2007, pp. 105-124.
- *padrão mais ou menos universal, (…) tendo como núcleo central as sete cores do espetro: vermelho, laranja, amarelo, verde, azul, anil e violeta*

From Biderman et al.

- Comparative study of two newpaper corpora from 1990-2000
- Words *azul* (blue), *vermelho* (red) e *encarnado*
- Noun and adjective, all forms
- Azul: PB 369, PP 673
- Vermelho: PB 452 PP 965
- Also: most frequent combinations (>5) with these words

Why is there less colour in BP?

- Several attempts
  - Fewer adjectives? Less modification in NPs?
  - Fewer"original colour" expressions?
  - Fewer genres involving colour in AC/DC?
  - Portuguese outliers such as political *laranja* and *capacete azul*?
  - Missing Brazilian colours?
- Other kinds of explanations
  - More coloured society – less attention to colour?
  - Colouring comes indirectly from reference to more coloured things?

Some (counter?)intuitive data

- If the sky is always blue, it is redundant to mention it
- If there is only one kind of feijão…
  - feijão branco, feijão verde, feijão encarnado (PT)
- The rarest eye colour is the most mentioned
  - PT: olho COLOUR: 252 azul:122, verde:36, … castanho:8
  - BR: olho COLOUR:161 azul:84, verde:35, … castanho:9

Further enquiry

- What are the N ADJ(colour) most related terms
  - CHAVE-BR: pasta, sinal, cabelo, camisa, olho, homem, movimento…
  - CHAVE-PT: luz, bandeira, espaço, vinho, cabelo, olho, homem…
  - But: *luz verde* (or *sinal verde*) is also metaphorical…and *bandeira azul* and *espaço verde* are technical
  - And *pasta cor-de-rosa* was topical
- What are the most common colour adjectives for sky?
  - BR (1317): Céu azul: 39, cinzento:2
  - PT (2537): Céu azul:41, cinzento:15, negro:6

Further enquiry

- What are the most common colour adjectives for sea?
  - PT azul 6, laranja 3, cor-de-rosa 1, branco 1, …
  - BR azul 9, verde 2, salino-cinza 1, …
- What are the most common colours for houses?
  - BR: multicolorido 1 verde-amarelo 1 transparente 1 vermelho 1
  - PT: amarelo 5 azul 3 verde 2 negro 2 vermelho 2 cor-de-laranja 1 cinzento 1 castanho 1 branco 1

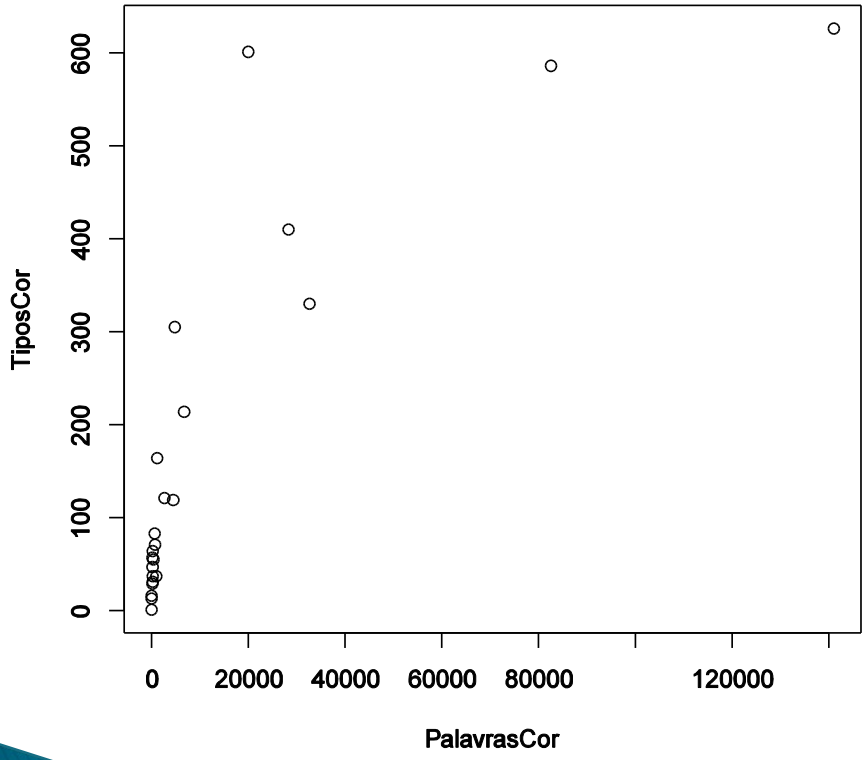
Colour distribution in CONDIV

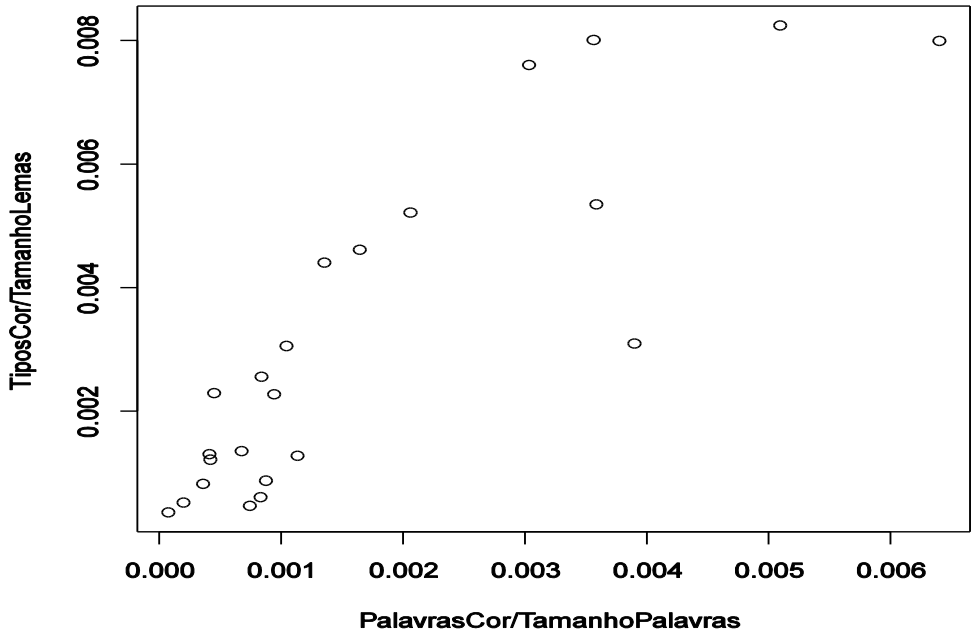
- Procura: **[sema="cor.*"]**.
  Distribuição de **sema**
  Corpo: CONDIVport 6.4 **20001** casos.
- **Distribuição:** Houve **8** valores diferentes de **sema**.
  - cor 15145
  - cor:equipa 2888
  - cor:original 1091
  - cor:humana 463
  - cor:ausência 299
  - cor:raça 80
  - cor_naomaduro 18
  - cor:vinho 17

Colour distribution in CHAVE

- Preliminary data, not revised: **82571** casos
- cor 61985
- cor:original 5649
- cor:raça 4143
- cor:humana 3662
- cor:equipa 2911
- cor:ausência 2134
- cor:política 1266
- cor:vinho 813
- cor_naomaduro 8

Corpora in terms of tokens and types





Replicability issues

- What is one counting? Tokens or instances of a concept?
  - Not all cases of *azul* concern colour
  - Not all cases of azul colour use the word *azul*
- Kinds of data
  - Forms
  - Lemmas: branco, branca
  - lemma-POS: brancoN, brancoADJ, brancaN
  - Group/category /profile
- Relative to
  - Corpus/no. of adjectives/sentence/phrase

The detailed study of language varieties… with the AC/DC cluster?

- Three moments: what is the material and how is it marked up?
  - Variety (**country**, province, social class, age, …)
  - Time of publication (decade, year, semester, day, …), time of writing
  - Genre, register, publication channel, author, …
  - Original/translated (from…)/transcribed
  - Revised at all?
  - Coherent or discontinuous?
- How comparable it is? How do intra-variety and inter-variety correlate?
  - Corpus homogeneity, corpus signature, or maximum quantity as the ideal good?

Support for formal variational linguistics

- Inspired by the Quantitative Lexicology and Variational Linguistics group http://wwwling.arts.kuleuven.be/qlvl/ at the Catholic University at Leuwen, and its Portuguese counterpart, CONDIVport, who developed a set of onomasiological profiles for the themes of football and fashion (health is underway)
- Linguateca did the same for colour, and revised annotation in context
- Both fashion and colour profiles were reused and improved and all AC/DC corpora were automatically annotated with them

Profiling…

- Profile names (fashion): ***blusa*** or ***blusão*** or ***calças curtas***
- Profile names (colours): ***vermelho*** or ***branco*** or ***creme***
- ***blusão***: *blazer*, *blusão*, *camurça*, *casaco de pele*, *colete*, etc.
- ***calças curtas***: *bermudas*, *calças à corsário*, *calças ¾* , *calções*, *shorts*, etc.
- ***vermelho***: *cor de carmim, cor de cereja, cor de chama, cor de colorau, cor de fogo alaranjado, cor de lagosta, cor de lagosta de viveiro, cor de morango, cor de morango esborrachado, encarniçado*, *escarlate*, *grená*, *magenta*, *ruborizar-se*, *rubro*, *vermelho-Benfica*, *vermelho-bordeaux*, etc.
- ***creme***: *aperolada*, *bege*, *bege África*, *bege-areia*, *marfim*, *cor de pele*, etc.

Comparing profile-based measures (Geeraerts & Grondelaers, 99)

$$A_{K,Z}(Y)=\Sigma_{i=1}^{n} F_{Z,Y}(x_i).W_{x_i}$$

- $A_{K,Z}(Y)$ is the ratio of terms with a feature K in the onomasiological profile for concept Z in dataset Y
- K= set of terms with a particular feature (for example FRENCH)
- Z= concept (for example VERDE, or VEST or BLUSÃO)
- $F_{Z,Y}$ relative frequency of x for concept Z in Y

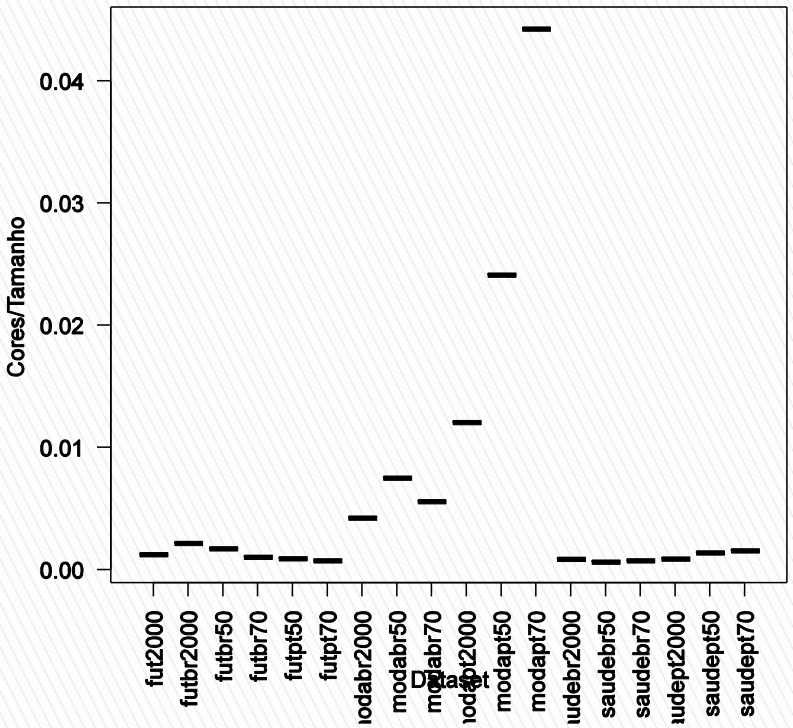
$$A_K(Y)=1/n* \Sigma_{i=1}^{n} A_{K,Zi}(Y)$$

- $A_K(Y)$ is the global proportion of the subset K in datase Y
- Comparing values of **relevant features** for different "datasets" (decades, varieties) convergence or divergence can be investigated

Discussion

- Can we apply this profiling to colours, assuming that they are different ways to describe the "same" meaning?
- We are entering the realm of properties, not objects… and properties in natural language are well known to be context dependent…
- We are not distinguishing formal vs. conceptual variation (*verde escuro*, *verde claro*)
- We are not distinguishig topical vs. non-topical uses of colour (expressions)

Comparing the various CONDIV subsets



Relative frequency of colour words

- fashion
- health
- football

- Decades
  - 50s
  - 70s
  - 2000s

Ideas for the future

- Remove (or dowtone) topical colour expressions automatically
  - Following Katz's model of "keywordness"
- Identify domain-specific terminological expressions with colour
- Check which colour features are most discriminating as
  - Genre identifiers
  - Variety identifiers: marron/castanho, 0/encarnado
- Produce a set of CONDIV-comparable corpora from the AC/DC cluster

To conclude

- Work in progress
- No point in presenting data based on non-revised corpora yet
- Linguateca's semantically annotated corpora aim for full coverage, not an automatically error-prone output
  - comparison with non-revised corpora will be provided soon (program being developed by Cristina Mota)