Linguateca's infrastructure for Portuguese... and how it allows the detailed study of language varieties

Diana Santos

A map of the talk

- Brief introduction of *Linguateca*
  - An infrastructure for Portuguese language technology
  - Short history
- The linguistic analysis of running text
  - Corpus projects for Portuguese
  - Three *Linguateca* projects: AC/DC, Floresta Sintáctica, and CorTrad
- Studying variation and varieties with the AC/DC cluster
  - Data
  - Formal variational linguistics support
  - New capabilities

Never heard about *Linguateca*?

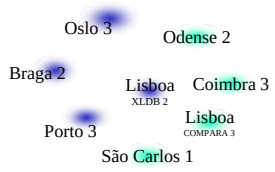
- It is a government funded initiative to significantly raise the quality and availability of resources for the **computational processing of Portuguese**
- After an initial plan for discussion by the community (white paper) a network was launched, headed by a small group (Linguateca's Oslo node) at SINTEF ICT (formerly SINTEF Tele og Data)
- This network has had as main goal to guarantee that
  - Information was provided and gathered at one place on the Web
  - Resources were made public, maintained, and further developed in connection with the scientific community
  - Evaluation initiatives were launched

Linguateca, a project for Portuguese

- A distributed resource center for Portuguese language technology

IRE model
- **I**nformation
- **R**esources
- **E**valuation

www.linguateca.pt



Linguateca highlights, www.linguateca.pt

- > 2000 links More than 7,000,000 visits to the Web site
- AC/DC, CETEMPúblico, COMPARA … Considerable resources for processing the Portuguese language
- *Morfolimpíadas* The first evaluation contest for Portuguese, followed by CLEF and HAREM

- Public resources
- Foster research and collaboration
- Formal measuring and comparison
- One language, many cultures
- Cooperation using the Internet
- Do not adapt applications from English

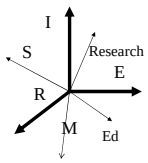
Linguateca's premises: not a research project

- a project whose aim is to considerably improve the conditions of the community who deals with the computational processing of the Portuguese language
- Is processing of Portuguese = NLP specialized to Portuguese? **NO**
- Does one build a community just by financing individual research projects? **NO**

- One has to **build a research infrastructure** and actively foster collaboration and joint evaluation

The IRE model and its evolution

- First: Information, Resources and Evaluation
- But then
  - (resource) Maintenance:
  - Support
  - Research (PhDs)
  - Education



A document to discuss the future of the area

- Main points: in 1998
  - There was hardly anything publicly available
  - People were alone doing the same things without knowledge of each other
  - No evaluation whatsoever
- Main need: an umbrella service
  - Maintaining and making resources available cannot be considered research
  - The sharing spirit for a common goal: open source philosophy
  - No separation of commercial/industrial and academic venues

At this moment, Linguateca is or has (produced)...

- Probably the largest repository on one language (computational processing) in the world (on the Web): kept at FCCN premises
- Well-known in the national communities (Portugal and Brazil) and in the international community (?)
- A set of reusable tools and resources that can be put to use by other researchers
- A set of studies on Portuguese and Portuguese processing (IR, GIR, MT, automatic terminology extraction, QA)
- A set of documents that enrich the area and can be used pedagogically
- A sizeable group of people trained in this area, a lot of others with some exposure to these activities through contact

Linguateca's achievements

- A lot of publicly available resources
- Several evaluation contests which advanced the state of the art
- Information, dissemination, gathering of relevant data and a team who answers
- The first evaluation contest for Portuguese
- The first treebank for Portuguese
- The first Web-based corpus service for Portuguese
- The first QA system for Portuguese
- The largest revised and annotated parallel corpus in the world
- The first national Web snapshot available

International impact

- Resources created by Linguateca available from the (Pennsylvania-based) Linguistic Data Consortium (LDC)
- Portuguese as one of the major languages in CLEF (more than 100 research groups worldwide participate in the largest evaluation forum for European languages and crosslingual information retrieval)
  - Linguateca belongs to the steering committee
  - Innovative pilots have been suggested by Linguateca, who has helped shaping the future
- The Portuguese treebank has often been used by third parties as example or resource in international venues, such as CoNLL or LREC
- According to Bernardo Magnini, Linguateca was the main inspiration for EVALITA, evaluation for Italian

Evaluation contests (*avaliação conjunta*)

Model: DARPA and NIST eval. cont.

- Jointly agree on a task and discuss the details together
- Create an evaluation setup
  - measures
  - resources
  - procedure
- Compare the performance of the several systems and get a state of the art
- Make public both resources, programs and systems' outputs for
  - external validation
  - research on both the task and the evaluation methodology
  - organization of future evaluation contests
  - training of newcomers

Linguistic analysis of running text

- Researchers on Portuguese needed support for computer-based empirical studies that were replicable and based on the same materials, available for extended periods of time, and that did not require physical access to specific premises
- Web-based services are the obvious answer, if they serve material that is curated and properly documented, and if they can be freely used
- AC/DC: providing access, making access possible
  - AC/DC cluster: a set of corpus projects, all inheriting from AC/DC, but with additional capabilities or features
  - Parallel corpora: COMPARA, CorTrad
  - Human revision: Floresta, COMPARA, ...

A brief history of Portuguese corpus linguistics

In the 1970s, oral corpora were collected
- *Português Fundamental* (inspired by the *Français Fondamental*)
- Projeto NURC (Labov-inspired)

Both in Portugal an Brazil, continuation of corpus studies
- VARSUL, Variação Lingüística Urbana do Sul do País (1982-)
- CRPC, Corpus de Referência do Português Contemporâneo (1988-)

In the 1990s, due to better computer facilities, a renewal/revival
- 1994 - CIPM, Corpus informatizado do português medieval
- 1998 - Tycho Brahe, *Padrões rítmicos, domínios prosódicos …*
- Projecto Natura, INESC, Corpus NILC/São Carlos, ...

A brief history of Portuguese corpus linguistics (ct)

- Banco de português (199x-)
- CORDIAL-SIN...DUPLEX (1998-)
- Português Falado - Variedades Geográficas e Sociais (1995-97)

International projects involving Portuguese
- CHILDES
- ENPC
- Borba-Ramsay corpus, ECI
- PORTEXT (1988-?)
- VISL (1994-)
- MLCC Multilingual and Parallel Corpora, *Official Journal of the EC*

Portuguese corpora during Linguateca's lifetime

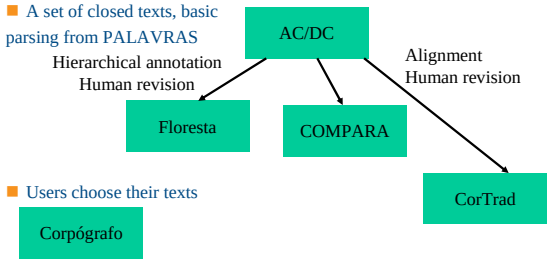
- Lácio-Web (2002-)
- C-ORAL-ROM (2001-2004)
- COMET (2005-)
- Corpus do português (2006-)
- etc.

- EuroParl
- Turigal
- JRC-Acquis

See also the ELC (*Encontros de linguística de corpus*) series in Brazil since 1999

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.)

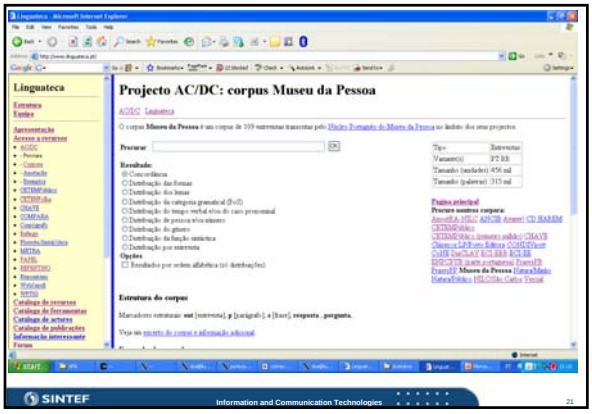
- Oral documents
  - Museu da Pessoa
  - ECI-EBR falado
  - Selva falado
- Technical
  - CorTradtec
  - ECI-EE
  - NILC/São Carlos tec
  - Selva Ciência
- Evaluation resources
  - CDHAREM
  - AmostRA
  - FrasesPP
- Email
  - Listas: ANCIB
  - SPAM: CoNE
- Web
  - Amazônia
- "Historical"
  - CETEMPúblico (primeiro milhão)
  - NatPublico

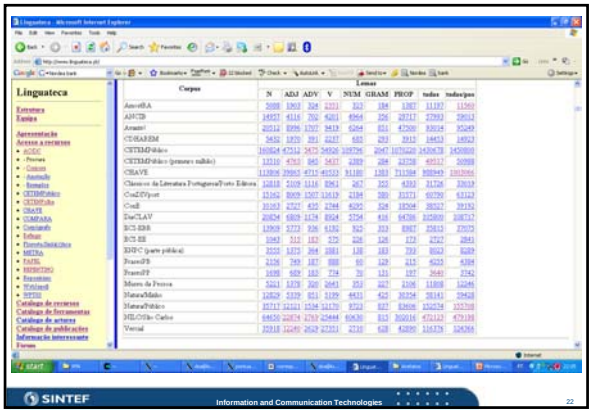
Adapted from Rocha (2007)

Brief description of AC/DC

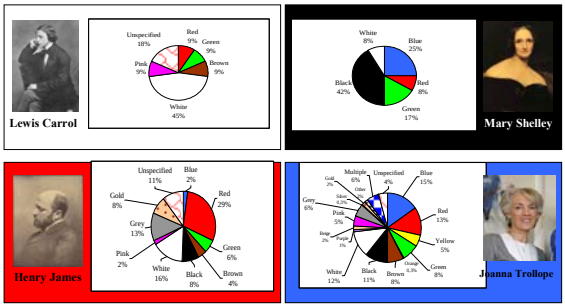
- **A**cesso a **C**orpora / **D**isponibilização de **C**orpora
- Ca. 20 different corpora
- Ca. 360 million words, 16 million sentences
- Portuguese and Brazilian varieties, a few other texts from others
- Different genres, mainly contemporary

- 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



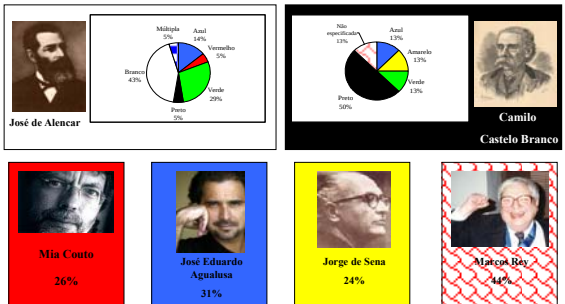


COMPARA: (EN) Author with highest % of colour:



Silva, Inácio & Santos (2008)

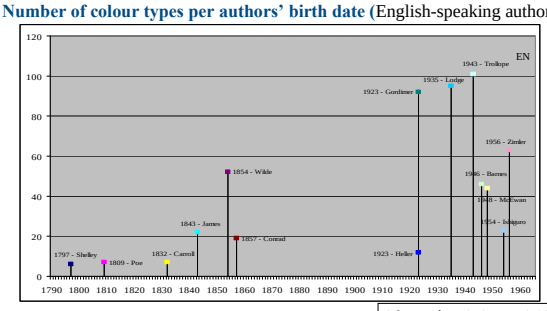
COMPARA: (PT) author with highest % of colour:



Silva, Inácio & Santos (2008)

COMPARA: Does colour quantity change with time?

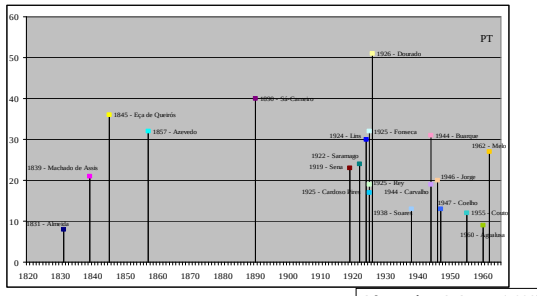
Number of colour types per authors' birth date (English-speaking authors)



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)

The Floresta Sintáctica treebank: history

- **2000** Formal cooperation between VISL and Linguateca started
- **2000-2001** Root planting: three linguistic scholarships at Odense, active preparation of tools and workflow at Oslo-Odense, launching of the basic resource and project philosophy (3-4 years work)
- **2002-2004** Partial work, incremental versions, stable but slow work
- **2005** Work on format validation at the Braga node

Sleeping forest...

- **2007-2008** New team, at Coimbra node: new material, new tools

Sleeping forest...

There is support and answer to questions, but not actual development

Floresta's "international" impact

- Used by Sabine Buchholz & Darren Green at their LREC 2006 article to illustrate treebanks' maintenance problems
- Used by Jason Balridge to infer a Portuguese grammar
- Used by CoNLL-X 2006, *ConLL-X shared task on multilingual dependency parsing* for Portuguese
- Integrated by Steven Bird in NLTK, *Natural Language Toolkit*, since September 2007
- Other explicit uses
  - John Hopkins University
  - Essex University
- Floresta anonymous downloads: (from our logs)

Some numbers from 2004

Clauses	21,931
Finite clauses	15,566
Infinite classues	5,602
Averbal clauses	763
Noun phrases*	43,096
Prepositional phrases*	32,210
Adjectival phrases*	1,780
Adverbial phrases*	833
Coordinated items	5,448
Trees	9,431

* phrase =*more than one word*

Lessons from Floresta

- Very hard to gather a community: much easier to create own treebanks
- Very hard to have any feedback from theoretical linguists
  - They were not invited in the first place!
  - Who is this German guy anyway?
  - Why do engineers talk about syntax?
- Very hard to have consensus on any subject whatsoever of linguistics:
  - What is a word? What is a phrase? What is a multiword expression?
  - What is an argument? What is an object? What is a phrase?
  - What is a head? What is a noun? What is a noun phrase?
- Ahead of our time? Impossible aspiration? Users will come in later?
  - Merge with current (independente) projects?

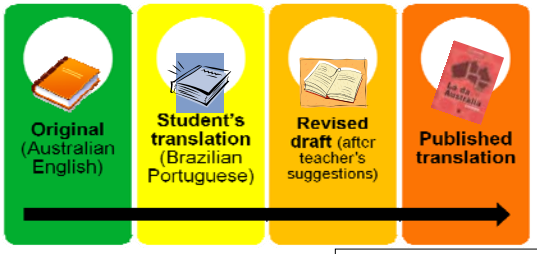
Brief presentation of CorTrad

- New material
  - Portuguese-to-English translation
  - Non-native translation vs. Native translation
  - Technical translation
- Multiversion
  - Study the changes from initial to published translation
  - Varieties might also be construed as rough translation to more idiomatic one
- Tailored for specific genres
  - Cookbook
  - Short scientific news

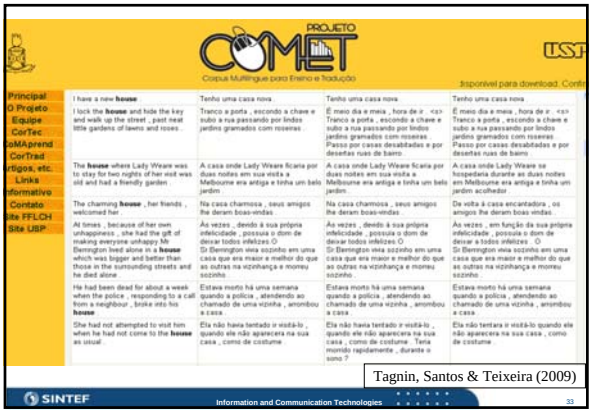
CorTrad is the parallel subcorpus of COMET, encoded with DISPARA, Linguateca's environment to make corpora available on the Web. A joint project of Univ. of São Paulo, Linguateca and NILC.

Literary CorTrad: Australian short stories

(*learner corpus)



Tagnin, Santos & Teixeira (2009)



Tagnin, Santos & Teixeira (2009)

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, created by Augusto Soares da Silva and his team at the Catholic Univ. of Braga, and made available through AC/DC, we started to provide support for this kind of studies as a merge with our semantic annotation efforts
- CONDIVport 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 dataset Y

- Comparing values of **relevant features** for different "datasets" (decades, varieties) convergence or divergence can be investigated

Current data about AC/DC profiles

- Number of different colour terms (lemmas) in the set of all corpora: **1672**, in 23 profiles (colour groups)
  (not counting proper names or other cases deriving from colour)
- Number of different clothing terms in the set of all corpora: **318**, in 28 profiles

	Colour tokens	Colour types	CT per 10^4
Condiv	20,380	547	47.5
CHAVE	85,506	526	5.47
CLASSLPPE	3,214	145	49.9
museudapessoa	167	24	20.0

Future capabilities in AC/DC

- Search helped/enhanced by semantic relations (synonyms, hypernyms, antonyms...) and syntactic relations (adj-past participle, clitics, contractions, nominalizations, ...)
- Reuse of complex queries
  - previous queries listed and explained (documentation effort, enriched FAQ)
  - programming of macros to make them more compact
- Automatic contrast between two different searches
  - Parallel results
  - Similarities
  - Contrasts

A preview of what may come...

- Compare X with Y according to...
  - Pure frequency
  - Distribution of lemmas
  - Distribution of passive/active, tense...
  - Presence of postmodifiers
  - Kind of subject, or of verb...
  - ...
- X and Y can be lexical items, or constructions, or any search whatsoever, restricted to whatever subsets
  - Compare ADJ N with N ADJ
  - Compare SEE pron VGER with SEE THAT finite
  - Compare *castanho* (PT) with *marrom* (BR), or in translation vs. original

Feedback highly appreciated!

We are open to

- Comments
- Questions
- Doubts
- Suggestions
- Ideas
- Critical remarks
- Cooperation proposals