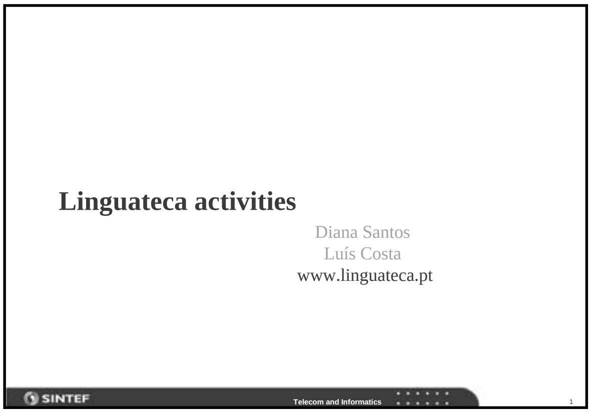
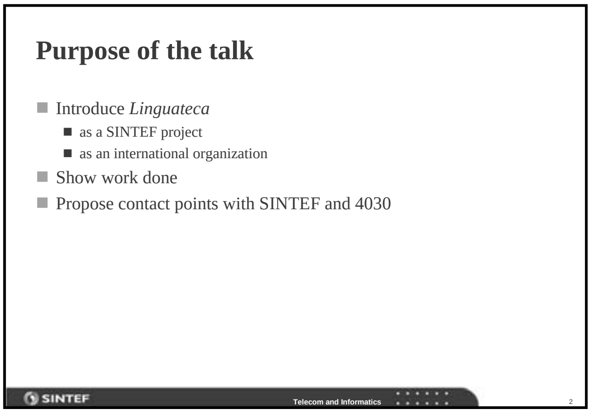


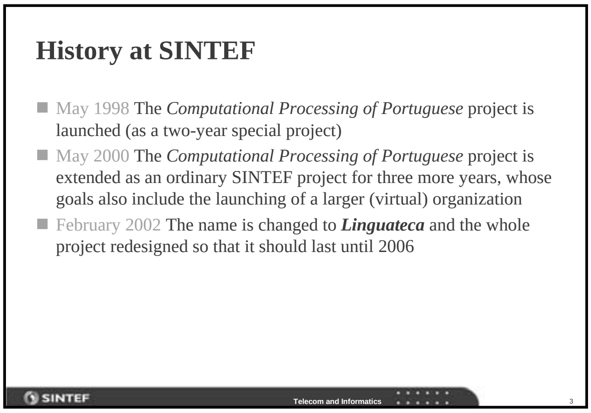
Linguateca activities

Diana Santos
Luís Costa
www.linguateca.pt



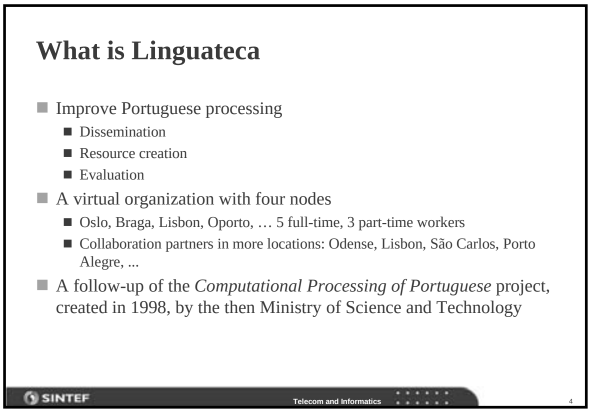
Purpose of the talk

- Introduce *Linguateca*
  - as a SINTEF project
  - as an international organization
- Show work done
- Propose contact points with SINTEF and 4030



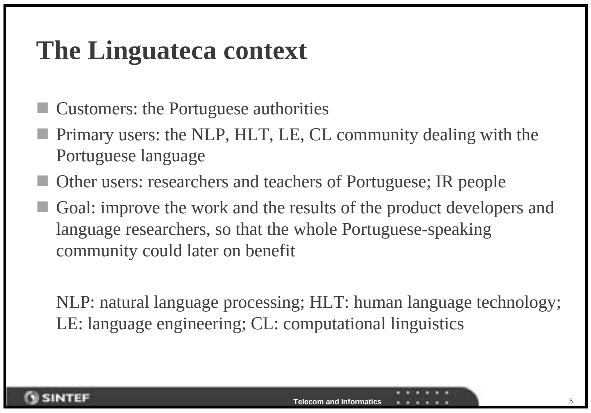
History at SINTEF

- May 1998 The *Computational Processing of Portuguese* project is launched (as a two-year special project)
- May 2000 The *Computational Processing of Portuguese* project is extended as an ordinary SINTEF project for three more years, whose goals also include the launching of a larger (virtual) organization
- February 2002 The name is changed to ***Linguateca*** and the whole project redesigned so that it should last until 2006



What is Linguateca

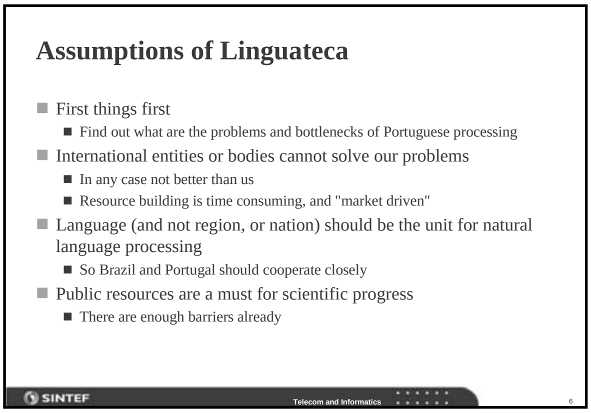
- Improve Portuguese processing
  - Dissemination
  - Resource creation
  - Evaluation
- A virtual organization with four nodes
  - Oslo, Braga, Lisbon, Oporto, … 5 full-time, 3 part-time workers
  - Collaboration partners in more locations: Odense, Lisbon, São Carlos, Porto Alegre, ...
- A follow-up of the *Computational Processing of Portuguese* project, created in 1998, by the then Ministry of Science and Technology



The Linguateca context

- Customers: the Portuguese authorities
- Primary users: the NLP, HLT, LE, CL community dealing with the Portuguese language
- Other users: researchers and teachers of Portuguese; IR people
- Goal: improve the work and the results of the product developers and language researchers, so that the whole Portuguese-speaking community could later on benefit

NLP: natural language processing; HLT: human language technology; LE: language engineering; CL: computational linguistics



Assumptions of Linguateca

- First things first
  - Find out what are the problems and bottlenecks of Portuguese processing
- International entities or bodies cannot solve our problems
  - In any case not better than us
  - Resource building is time consuming, and "market driven"
- Language (and not region, or nation) should be the unit for natural language processing
  - So Brazil and Portugal should cooperate closely
- Public resources are a must for scientific progress
  - There are enough barriers already

Linguateca activities

- Dissemination of information and resources on Portuguese processing
  - Web catalogue with a dedicated search engine
  - Forum and a contact service
- Creation of publically available language resources
  - Making the available resources more available: Web services
  - Creating new ones: both Web and physical access
- Promotion of joint evaluation using the evaluation contest or evaluation campaign model
  - Web site and discussion list [avalia]
  - Organization of a workshop (June 2002) and a conference (AVALON' 2003)
  - Organization of the first evaluation contest for Portuguese: *Morfolimpíadas*

Dissemination: some numbers

- size of site
  - 1,047 Web pages
  - 1,392 resource links
  - 643 own documentation
  - 723 publication entries
- size of audience (1st May 2003)
  - number of visits: 926,887
  - number of queries to our on-line services: 50,954
- size of recognition
  - 685 Web pointers to us
- published papers or reports, and other presentations
  - 24 (+2) in Portuguese; 15 (+4) in English

Creation of language resources

- copyright clearing
- creation
- programming resource specific tools
- testing
- version dealing
- producing information and documentation
- evaluation
  - does it meet the goals?
  - is it being correctly used?
- giving support

Resource creation and dissemination

- AC/DC: querying a variety of (annotated) corpora
  - developed outside (rights obtained), or in-house
- COMPARA: querying English-Portuguese
- CETEMPúblico and CETENFolha: large amounts of newspaper language, divided in extracts and scrambled
- *Floresta Sintá(c)tica*: manually revised syntactically analysed text
- Web services
  - AC/DC service collection
  - DISPARA
  - *Águia*
  - AneLL (Lisbon): morphosyntactic tagging of private texts
  - GC (Oporto): comparable corpora environment (English-Portuguese)

COMPARA

- On-going collaboration with Ana Frankenberg-Garcia
- Text team (Lisbon) and engineering team (Oslo): email communication; clearly defined workflow, with at least six steps for each text pair
- A general Web system for parallel corpora, DISPARA, evolved
- Currently 29 text pairs; 36 in the processing queue
- 12,500 queries since May 2000 from all over the world

http://www.linguateca.pt/COMPARA

Floresta Sintá(c)tica

The first treebank for Portuguese

- Collaboration with Eckhard Bick and the VISL project (Odense)
- Main activities: October 2000 to December 2001; a few things added afterwards
- Workflow: a complex process with several revision steps and three different automatic modules (a parser, a tree transducer and a CQP converter)
- Tools: Pica-Pau, a tree editor; Águia, a Web interface
- Resource: 1,500 trees (ca. 35,000 words) in phrase structure format and in CG dependency format, both Web searchable and downloadable
- Sub-projects: inter-annotator test; sentence separation evaluation; streamlined revision using Águia; use as golden standard in Morfolimpíadas
- Status: waiting for renovation; discussion in Avalon’2003

http://www.linguateca.pt/Floresta/

Evaluation

- The most challenging task
- History:
  - Tutorial on evaluation of NLP systems in Atibaia (Brazil), 2000
  - Some papers on resource and problem evaluation (2001, 2002)
  - Movement with a Web site and a dedicated mailing-list in 2002
  - Preparatory encounter dedicated to “joint evaluation” June 2002
  - Morfolimpíadas
    - Trial in September 2002 - March 2003
    - Contest May-June 2003
  - Avalon’2003
    - Named entity recognition
    - Portuguese IR
    - Machine translation and alignment
    - Some syntax evaluation

Morfolimpíadas: cooperatively evaluating morphological analysers for Portuguese

- Evaluation contest paradigm
  - Importance for science and for community building
  - Shared task, consensual result, objective measures, knowledgeable organization
- Why morphology
  - Mildly inflected language (70 verb forms)
  - Simple and well defined (?) problem, no infinite set of members
  - Traditionally the first module in a set of NLP tools
  - The task for which there was greater interest
- Goals
  - Exemplify the paradigm with a relatively short schedule
  - Assess the state of the art in morphology (also looking at tokenization)
  - Measure the problem

1.^{as} Morfolimpíadas: overview

- Seven participating systems, out of 16-20 out there
  - 3 Portugal 2 Brazil 2 Int
  - 5 “real” morphological analysers, 1 spellchecker and 1 stemmer
- Organization: Linguateca Oslo (+Oporto+contractors)
- Setup:
  - Registration, providing some data
  - Ran their system over 80,000 running text words, in three different formats
  - Processing:

uts.SYSTEM.def.preze.ze.hi.gr.un.le

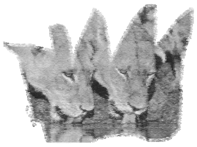
Zebras: transform into an internal format

- Every system (with a wildly different output format) is turned into “zebraic” format
- Every zebra output is apparently similar but intriguingly different
- Zebras may still require hienas to deal with complex issues (clitics and contractions)
- Zebra programming requires a full understanding of the high and low level details of the systems (underlying linguistic conception, tokenization behaviour)

Further processing

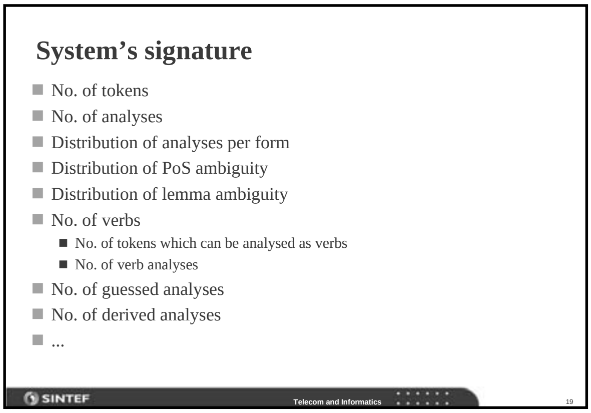
- Grammatical analyses are turned into one analysis named GRAM
- Some sets of always ambiguous interpretations in the verbal paradigm are turned into one
  - first and third person singular of some tenses
  - personal and impersonal infinitive
  - third person plural of Perfeito and Mais que perfeito
- Numbers are dealt with in a simple form
- Punctuation marks and proper names are handled to yield a hopefully more similar output
- Tokenization problems are dealt with to some extent

Leoas: tearing files to pieces



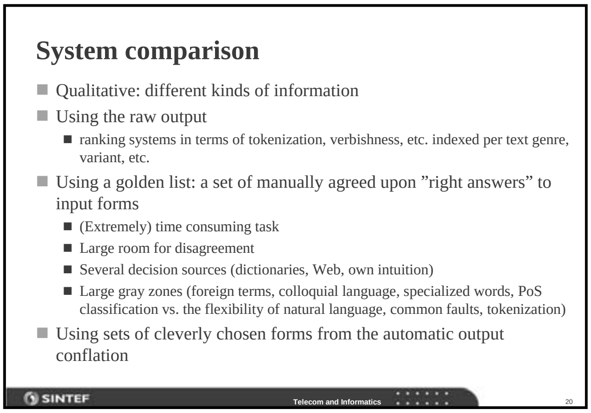
- Distribution by text
- Distribution by variant
- Distribution by genre
- Distribution by medium

- Rationale:
  Is system performance correlated with type of text? Variant?



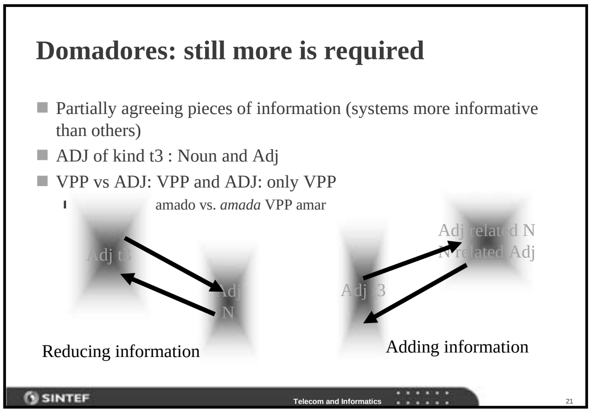
System’s signature

- No. of tokens
- No. of analyses
- Distribution of analyses per form
- Distribution of PoS ambiguity
- Distribution of lemma ambiguity
- No. of verbs
  - No. of tokens which can be analysed as verbs
  - No. of verb analyses
- No. of guessed analyses
- No. of derived analyses
- ...



System comparison

- Qualitative: different kinds of information
- Using the raw output
  - ranking systems in terms of tokenization, verbishness, etc. indexed per text genre, variant, etc.
- Using a golden list: a set of manually agreed upon ”right answers” to input forms
  - (Extremely) time consuming task
  - Large room for disagreement
  - Several decision sources (dictionaries, Web, own intuition)
  - Large gray zones (foreign terms, colloquial language, specialized words, PoS classification vs. the flexibility of natural language, common faults, tokenization)
- Using sets of cleverly chosen forms from the automatic output conflation



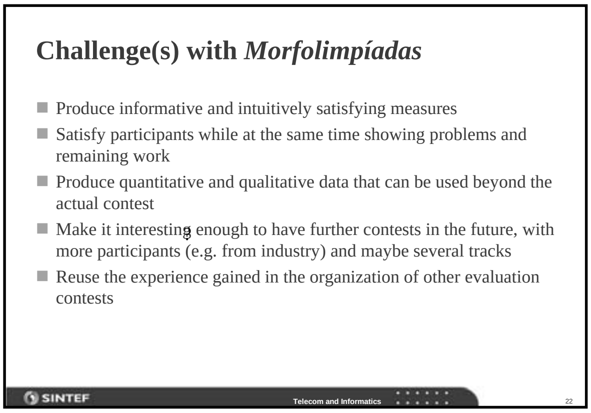
Domadores: still more is required

- Partially agreeing pieces of information (systems more informative than others)
- ADJ of kind t3 : Noun and Adj
- VPP vs ADJ: VPP and ADJ: only VPP
  - amado vs. *amada* VPP amar

Adj t3
Adj
N
Adj t3
Adj related N
N related Adj

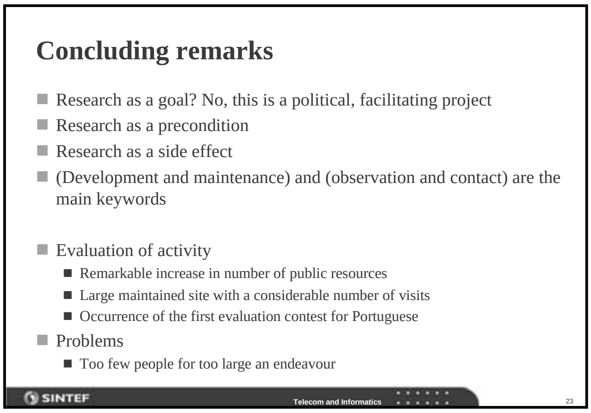
Reducing information

Adding information



Challenge(s) with *Morfolimpíadas*

- Produce informative and intuitively satisfying measures
- Satisfy participants while at the same time showing problems and remaining work
- Produce quantitative and qualitative data that can be used beyond the actual contest
- Make it interesting enough to have further contests in the future, with more participants (e.g. from industry) and maybe several tracks
- Reuse the experience gained in the organization of other evaluation contests



Concluding remarks

- Research as a goal? No, this is a political, facilitating project
- Research as a precondition
- Research as a side effect
- (Development and maintenance) and (observation and contact) are the main keywords

- Evaluation of activity
  - Remarkable increase in number of public resources
  - Large maintained site with a considerable number of visits
  - Occurrence of the first evaluation contest for Portuguese
- Problems
  - Too few people for too large an endeavour