

# The *Linguateca* experience: What can be reused? Or, what have we learned? Diana Santos

## A map of the talk

- Brief introduction about *Linguateca*
  - Starting point
  - End
- Achievements
- Failures
- 
- Future
  - Promising alleys
  - Blind alleys

## 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 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

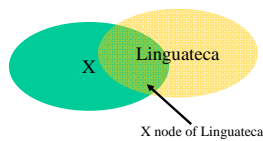
### IRE model

- Information
  - Resources
  - Evaluation
- [www.linguateca.pt](http://www.linguateca.pt)



## Organization details

- Organization nucleus at SINTEF ICT
- All other nodes or branches are in scientific institutions such as universities or public R&D bodies, in order to cross-fertilize the activity of that group with the rest of the network and the general public



## *Linguateca* highlights, [www.linguateca.pt](http://www.linguateca.pt)

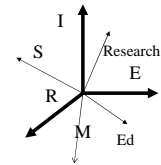
- > 2000 links More than 7,000,000 visits to the Web site
- [AC/DC](#), [CETEMPúblico](#), [COMPARA](#) ... Considerable resources for processing the Portuguese language
- *Morfolimpíadas* The first evaluation contest for Portuguese, followed by CLEF and HAREM
- Public resources
- Foster research and 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



## In the beginning

- There was a little project at SINTEF (1998-2000), Diana Santos and Signe Oksefjell
- which produced a white paper  
Diana Santos. "Computational processing of Portuguese: working memo". 1999.  
written for a general discussion in Portugal of what to do to considerably improve this area
- and started what later on was called *Linguateca*
  - creating a portal for CPP
  - starting corpora services on the Web

## 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

## The end (2008)

- Probably the largest repository on one language (computational processing) in the world (on the Web): it is going to be kept by FCCN
- 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

## International impact of the involved researchers

- Several invitations for lectures abroad related to resources or themes developed under the scope of Linguateca
- Participation in international summer schools or courses as teachers
- Steady invitation for program committees and conference venues

## Administrative details

Timeframe	Total	SINTEF	
■ May 1998 - May 2000:	270 k€	270 k€	
■ May 2000 - May 2003:	1.4 M€	970 k€	
■ May 2003 - Dec. 2006:	1.8 M€	1 039 k€	
■ Dec. 2006 - Dec. 2008:	1.7 M€	900 k€	
■ 10 years and 7 months	4.8 M€	3.2 M€	(66%)
■ In NOK	38.4 MNOK	25.6 MNOK	

## Evaluation contests (*avaliação conjunta*)

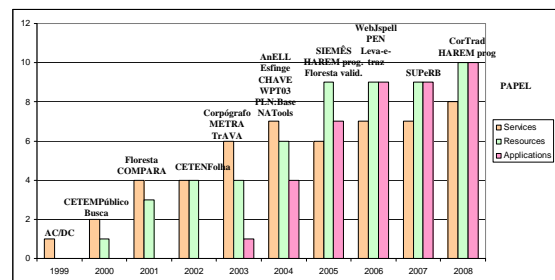
Model: DARPA and NIST eval. cont.

- 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

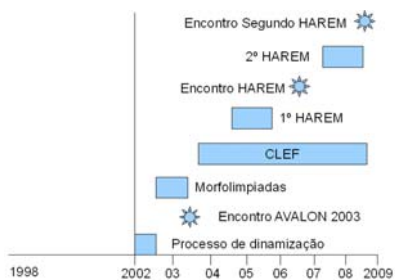
## Further advantages of an evaluation contest

- Agree on details that generally make individual evaluation measures incommensurable
- Raise awareness about a particular task, its problems and solutions: community building
  - several new systems were born with HAREM
- Produce a wealth of documentation that otherwise would never have been produced
  - cf. HAREM guidelines; cf. the wide discussion of particular morphological problems and solutions; the discussion around QA systems in CLEF
- Can provide baselines and resources (systems, gazetteers) for other work

## Resources, services and applications created in the scope of Linguateca

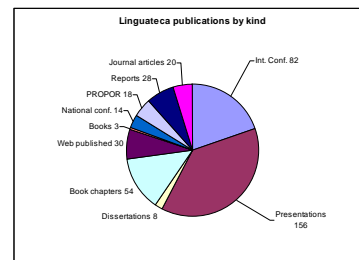


## Evaluation initiatives

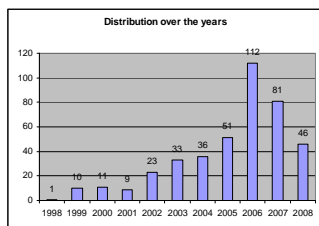


## Publications and presentations 1998-2008

- Total: 413
- Journals: 20
- Conference papers: 100+14
- Book chapters: 54
- Books: 3
- Theses: 8
- Presentations: 156
- Reports: 58



## Distribution over the years



## Linguatca's failures

- People (re)use without citing or acknowledging
- People get money/projects to do the same with total impunity
- People compare their results unfairly without participating in our evaluation contests
- People still prefer to participate in "international" conferences/evaluation contests although they are much less interesting in scientific terms
- People still prefer to publish in (Portuguese-speaking program committees) Springer and/or in (bad) English

One may say this is outside our competence, but still it goes counter our basic premises and choices

## Position in Norway/Nordic countries

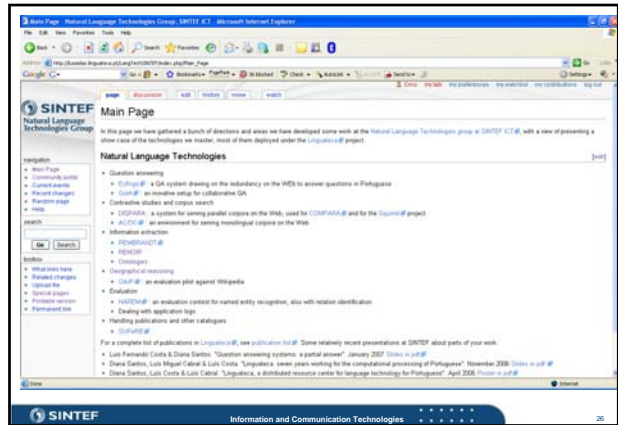
- Ignored!!!
- People in charge of reports on "norsk språkteknologi" disregarding *Linguatca* because it was concerned with Portuguese
- No interest whatsoever in free courses
  - Evaluation in ESSLLI
  - Using Portuguese corpora for teaching Portuguese
- Attempt to use the Portuguese treebank in a Nordic context (Swedish and then Nordic) was a failure
  - They preferred people with no experience in treebanks but with knowledge of Norwegian
- The few CL partners that show some interest always give the excuse: SINTEF prices are too high

## Looking forward: a 20 years' Janus?

- From building and leading an international network...
  - ... what can this small group now do?
    - Do "the same" for Norwegian
      - The same is obviously different
    - Use our competence to do different things
      - For Norwegian
      - For crosslingual/multilingual
      - For crossmodal
      - For R&D management

## Can Linguateca's experience help Norwegian?

- We have learned a lot with our work in Linguateca
  - Technical infrastructure
  - Communication in a highly distributed setup
  - How to organize evaluation contests
  - What does not pay, and what could have been done better, if there were coordination with project funding
- Since most of the work has been done at SINTEF ICT, it is obvious that we should try to reuse it for Norwegian, given SINTEF's motto *Teknologi for en bedre samfunn*
- We do not need to start from scratch, we can reuse significant parts of our work for Portuguese for Norwegian



## Language engineering at SINTEF

- Question answering
- Ontologies
- Geographical reasoning
- Contrastive studies
- Information extraction (NER, etc.)
- Corpus search



Publication management  
Log analysis

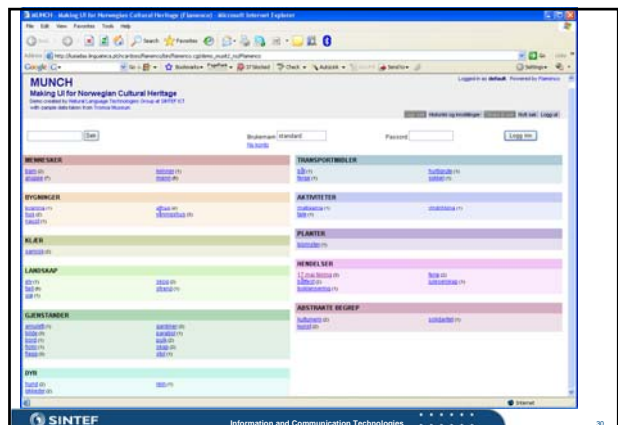
We believe all these pieces will help us to address tasks that are not purely NLP (Norwegian language processing) but also cross-media and cross-application

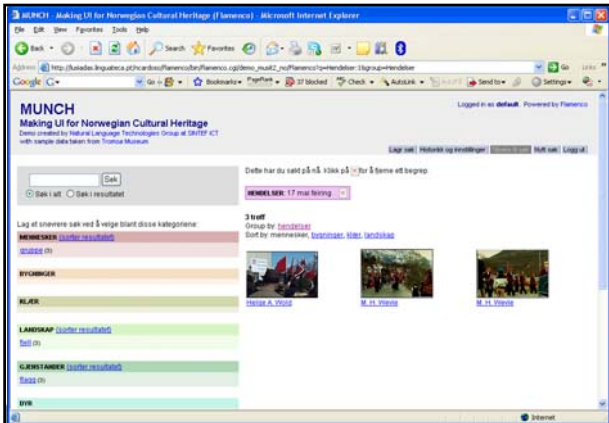
## PIADA in a Norwegian context

- We want to develop specific knowledge on images in Norwegian*
- The vocabulary of images and image search in Norway
  - Study of user behaviour: what do people ask for, what do they want to see?
  - Picture reasoning ontologies
  - By cooperating with commercial actors we will do something useful – not only for research purposes

## Language technology and museums

- A place where information extraction, NLP and images meet
- Help increase access to a wealth of ethnographic, historical and scientific material which is property of the (university) museums
- The idea is to create a user-friendly access to the data, with a faceted view,
  - massaging, analysing and clustering the descriptions
  - semantically tagging the categories involved
  - studying user queries and folksonomy-like views of the museum data
  - bridging ontologies
  - geographical reasoning





## Extending international evaluation to Norwegian

- Linguateca is in the steering committee of CLEF
- We have suggested two pilots who were accepted
  - QoLA in 2006 – given up due to Linguateca's interruption end 2006
  - GikiP in 2007 – took successfully place in 2008
- Use GikiP / QA / GeoCLEF as a way of forming a community interested in applications that deal (also) with Norwegian
- ...
- Basically, the most important argument is that further participants will come to CLEF

## What is GikiP?

- GikiP is a pilot evaluation task run in the CLEF evaluation contest.
- Task: *Find Wikipedia entries (i.e. articles) that answer a particular information need which requires geographical reasoning of some sort.*
- Scientific goal: Create synergies between geographic information retrieval (GIR) systems and question answering (QA) systems.
- Practical goal: Wouldn't it be good if we had systems that could mediate between us & Wikipedia, and answer our complex questions, no matter the language?

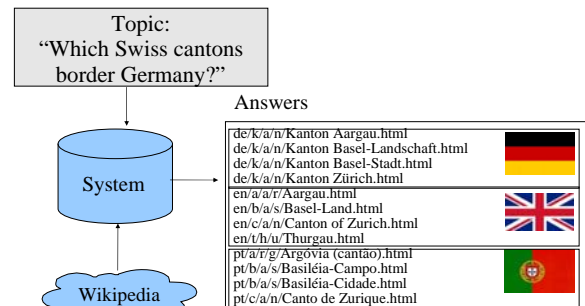
## Topic titles in GikiP 2008

ID	English topic title
GP1	Which waterfalls are used in the film "The Last of the Mohicans"?
GP2	Which Vienna circle members or visitors were born outside the Austria-Hungarian empire or Germany?
GP3	Portuguese rivers that flow through cities with more than 150,000 inhabitants
GP4	Which Swiss cantons border Germany?
GP5	Name all wars that occurred on Greek soil.
GP6	Which Australian mountains are higher than 2000 m?
GP7	African capitals with a population of two million inhabitants or more
GP8	Suspension bridges in Brazil
GP9	Composers of Renaissance music born in Germany
GP10	Polynesian islands with more than 5,000 inhabitants
GP11	Which plays of Shakespeare take place in an Italian setting?
GP12	Places where Goethe lived
GP13	Which navigable rivers in Afghanistan are longer than 1000 km?
GP14	Brazilian architects who designed buildings in Europe
GP15	French bridges which were in construction between 1980 and 1990

## GikiP's collection: Wikipedia

- Wikipedia is a great collection to work on:
- Truly multilingual (dozens of languages)
  - Spans several subjects, we can dare to say more interesting than newspaper subjects...
  - Documents are well written, reviewed and validated on its content
  - Rich content, structure and metadata that can be explored (categories, infoboxes, links)
  - Multimedia resource (it mixes text, images and may have links to sound quotes or videos)

## GikiP: one simple example



## The system should...

- ...understand what the topic really wants (a list of cities, rivers or mountains), and its restrictions (a given population/length/height threshold)
- ...reason over the Wikipedia collection and over the geographic domain (i.e., "does this river flows to the Atlantic Ocean?")
- ...return Wikipedia pages for the answers: not lists, not overview pages, just the answers.

SINTEF

Information and Communication Technologies

## Interesting issues (1)

- Names change, roles change!
- Topic: "African capitals..."



SINTEF

Information and Communication Technologies

## Interesting issues (2)

- Different languages, different meanings of geographic scope
- *Australia*: both a continent and a country in EN, but only a country in PT (continent: *Oceânia*)
- *The highest mountains of Australia...*



SINTEF

Information and Communication Technologies

## Interesting issues (3)

- Different languages, different information sources
- Ex: *African capitals with more than x habitants*

Wikipedia EN on "Harare":	Wikipedia PT on "Harare":	Wikipedia DE on "Harare":																														
<table border="1"> <tr><th>Harare</th></tr> <tr><td><b>Capital</b> Harare</td></tr> <tr><td><b>População</b> 1.800.000 habitantes</td></tr> <tr><td><b>Country</b> Zimbabwe</td></tr> <tr><td><b>Province</b> Harare</td></tr> <tr><td><b>Founded</b> 1930</td></tr> <tr><td><b>Incorporated (city)</b> 1955</td></tr> <tr><td><b>Government</b></td></tr> <tr><td><b>Mayor</b> Muchawab Mabumba</td></tr> <tr><td><b>Elevation</b> (ft) 1.490 m (4.889 ft)</td></tr> <tr><td><b>Population</b> (2006)</td></tr> <tr><td><b>City</b> 1.800.000</td></tr> <tr><td><b>Urban</b> 2.800.111</td></tr> <tr><td><b>metropolitan</b> unknown</td></tr> </table>	Harare	<b>Capital</b> Harare	<b>População</b> 1.800.000 habitantes	<b>Country</b> Zimbabwe	<b>Province</b> Harare	<b>Founded</b> 1930	<b>Incorporated (city)</b> 1955	<b>Government</b>	<b>Mayor</b> Muchawab Mabumba	<b>Elevation</b> (ft) 1.490 m (4.889 ft)	<b>Population</b> (2006)	<b>City</b> 1.800.000	<b>Urban</b> 2.800.111	<b>metropolitan</b> unknown	<table border="1"> <tr><th>Harare</th></tr> <tr><td><b>Capital</b> Harare</td></tr> <tr><td><b>População</b> 1.800.000 habitantes</td></tr> <tr><td><b>Country</b> Zimbabwe</td></tr> <tr><td><b>Area</b> 872 km²</td></tr> <tr><td><b>Densidade</b> 2.182,92 hab/km²</td></tr> <tr><td><b>Mapa</b></td></tr> </table>	Harare	<b>Capital</b> Harare	<b>População</b> 1.800.000 habitantes	<b>Country</b> Zimbabwe	<b>Area</b> 872 km²	<b>Densidade</b> 2.182,92 hab/km²	<b>Mapa</b>	<table border="1"> <tr><th>Wappen</th></tr> <tr><td></td></tr> <tr><th>Karte</th></tr> <tr><td></td></tr> <tr><th>Geographische Lage</th></tr> <tr><td><b>Höhe</b> 1.490 m ü. NN</td></tr> <tr><td><b>Fläche</b> 872 km²</td></tr> <tr><td><b>Einwohner</b> 1.800.000 (2006)</td></tr> <tr><td><b>Bevölkerungsdichte</b> 2.183 Einwohner/km²</td></tr> </table>	Wappen		Karte		Geographische Lage	<b>Höhe</b> 1.490 m ü. NN	<b>Fläche</b> 872 km²	<b>Einwohner</b> 1.800.000 (2006)	<b>Bevölkerungsdichte</b> 2.183 Einwohner/km²
Harare																																
<b>Capital</b> Harare																																
<b>População</b> 1.800.000 habitantes																																
<b>Country</b> Zimbabwe																																
<b>Province</b> Harare																																
<b>Founded</b> 1930																																
<b>Incorporated (city)</b> 1955																																
<b>Government</b>																																
<b>Mayor</b> Muchawab Mabumba																																
<b>Elevation</b> (ft) 1.490 m (4.889 ft)																																
<b>Population</b> (2006)																																
<b>City</b> 1.800.000																																
<b>Urban</b> 2.800.111																																
<b>metropolitan</b> unknown																																
Harare																																
<b>Capital</b> Harare																																
<b>População</b> 1.800.000 habitantes																																
<b>Country</b> Zimbabwe																																
<b>Area</b> 872 km²																																
<b>Densidade</b> 2.182,92 hab/km²																																
<b>Mapa</b>																																
Wappen																																
Karte																																
Geographische Lage																																
<b>Höhe</b> 1.490 m ü. NN																																
<b>Fläche</b> 872 km²																																
<b>Einwohner</b> 1.800.000 (2006)																																
<b>Bevölkerungsdichte</b> 2.183 Einwohner/km²																																

SINTEF

Information and Communication Technologies

## Interesting issues (4)

- Not all questions can be answered easily by a person!
- For example: "Name all wars that occurred on Greek soil"
  - There is no straightforward category in Wikipedia to start with.
  - Even if there was a "Greek War" category, does it really includes only wars taken on Greek soil, or all wars involving Greece?
  - Temporal issues: How was the Greek soil back then? Narrower or longer than today's Greek boundaries?

SINTEF

Information and Communication Technologies

## Interesting issues (5)

- Reasoning over the geographic domain
- Topic GP11: "Which plays of Shakespeare take place in an Italian setting?"

SINTEF

Information and Communication Technologies

## GikiP's future (1)

- Why not mix images and text?
- Example: "Name the countries that still have lynxes"



## GikiP's future (2)

More complex topics also balancing images and text

- "Portuguese cities founded before 1500 with rivers larger than 100 km and featuring a Moorish castle"
- "Which Swiss cantons have a lion on their flag?"
- "Find portraits of married women in the 18th century"

In other words, users express their needs clearly in their language; the systems must adapt to the user, and not the other way.

## Publications catalogue as a product for institutions

It is still not yet possible to get on the SINTEF web publications ordered by any field (author, date of publication, etc.)...

## SUPeRB is miles ahead!

- Since there is money to get with a correct and timely publication reporting, maybe one could use our experience instead of contacting "outside leverandører"
- Only people with knowledge of what scientific publication is about should be in charge of developing scientific publication management and reporting systems, not administrative or technical personnel
- Although dealing with publications is not mainly language technology, it does have some part of it: and can be extended in order to have much more!

## Examples of what extended SUPeRB could do

- How many SINTEF (9012) researchers published in a journal in 2005?
- How many publications of level 2 exist in SINTEF (9012)'s database?
- What institutions are co-authoring with SINTEF (9012) researchers?
- What are the most popular conferences for the KST group?
- Which publications have more than three authors from SINTEF?
- List the publications in conferences in the USA of authors X and Y.
- Draw the distribution of publications by SINTEF (9012) group.

## Summing up

- We have developed competence in language processing and information retrieval with the view of processing one language well
- We can do the same – use the same technologies and the same methodologies and similar underlying resources – for Norwegian, provided we cooperate with Norwegian linguists/language experts for the final details of the user interface
- But there is a lot of work that has to be done for Norwegian and we are trying to find a way to cooperate with other groups so that
  - we will not reinvent the wheel
  - but will be able to compete friendly for funding
- Crosslingual, multilingual or monolingual?

## The beginning ☺

- Back to work!
- *Technology for a better society*

## Acknowledgements

- Linguateca funding came from contract no. 339/1.3/C/NAC, project jointly funded by the Portuguese Government and the European Union, and administratively led by FCCN.

