THORNY ISSUES IN NLP RESEARCH: a personal view

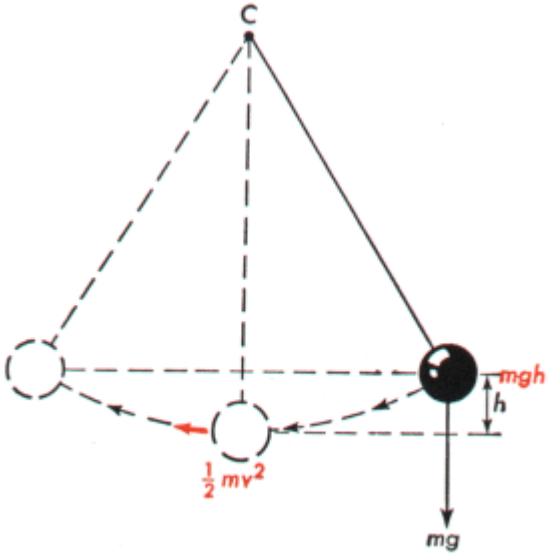
Diana Santos

A personal timeline

- 1987-1990: IBM-supported research
- 1991-1996: fishing for funding, only a PhD grant
- 1994: teaching NLP at IST
- 1997-1998: support at the Arts Faculty
- 1998-2010: “research” at SINTEF, within Linguateca
- 2010- : back to technical support, now in a HPC site

Reflections after 20++ years

- The binary side of NLP (?)
  - Interdisciplinarity, crossdisciplinarity
  - Rationalism vs. empiricism
  - Theory vs. practice
  - Old vs. new
  - Evaluation vs. value
- Academia vs. society
  - Funding models
  - Evaluation models



Interdisciplinaridade

- Uns vêm de Letras
- Outros de informática

- Mas o caminho a seguir é o mesmo, na área do processamento computacional da língua

- Não interessa onde começaram
- Professores dos 2 lados

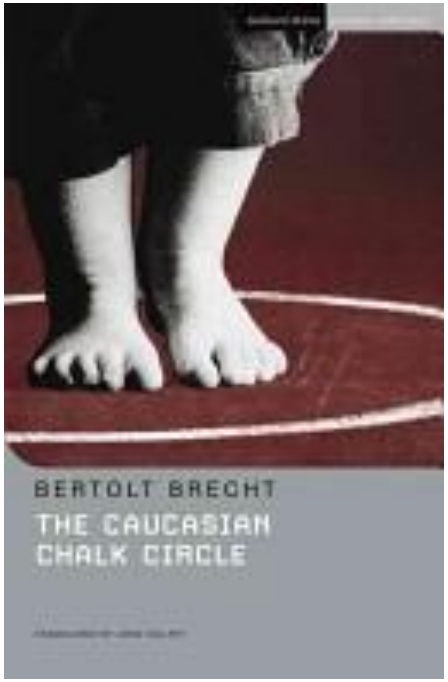
Primeira Escola de Verão da Linguateca

Blessing or fault?

- To do NLP you need to know as much linguistics as engineering/computer science
- You are not understood by either group of "core" practitioners, who tend to dismiss the "other area"
  - *Language is just an application area*
  - *Engineering methods are just helping tools*
- An interesting side effect: practictioners are often too lenient as far as their own core area is concerned

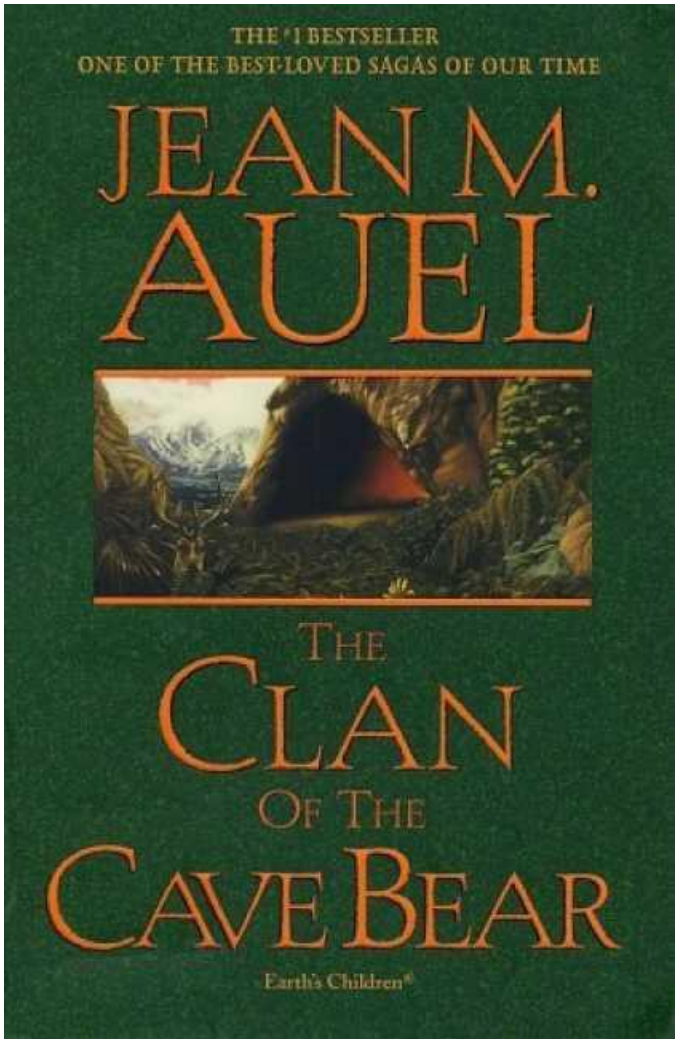
Examples of problems/mismatches

- A highly sophisticated computational system… to find out that to abstract/simplify text you can remove adverbs and adjectives
- A highly sophisticated linguistic analysis… to be employed to create a ML PoS tagger
- A highly complex linguistic problem, such as translation, dealt with by simple and inadequate computational procedures



Rationalism vs empiricism

- Who is right? Our understanding or the data?
- Is our understanding dependent on the language input? Or are we programmed to understand?
- How to link the potential ability to the actual performance? (Chicken and egg)
  - Google storing



Science **and?** technology

- Technology produces artifacts, tools
  - Language engineering produces artificial languages
  - But natural language is spoken by humans
- NLP technology should not produce another language, even though the actors are computers
  - Chemical industry produces new substances
  - But NLP technology can produce (new) texts
  - And can help humans deal with their own texts, cf. computer-assisted “everything”

Science… or engineering?

What can be wrong with this algorithm?

- *Take 86 pairs of Web pages that may be parallel texts, ask people to detect whether they are parallel or not, and then create a golden collection*

(Resnik & Smith, 2003)

What may be wrong with this system?

- *Right answer in 90% of the cases*

(Artstein & Poesio, 2008)

What can go wrong…

- … if one compares entries from lang1 -> English and Lang2 -> English to get a lang1-lang2 dictionary?

(Sjöbergh, 2005)

- … if one compares two documents based on the proper names they share?

(Friburg & Maurel, 2002)

- … if one tries to analyse Wikipedia pages based on the categories they belong to?

(Cardoso, 2009, p.c)

Too many categories?

- http://en.wikipedia.org/wiki/**Empire_State_Building**
- Categories: 1931 architecture | **Accidents involving fog** | Art Deco buildings in New York City | Fifth Avenue (Manhattan) | Former world's tallest buildings |National Historic Landmarks in New York City | Office buildings in New York City | National Register of Historic Places in Manhattan | Skyscrapers in New York City | Skyscrapers over 350 meters | Visitor attractions in New York City | Tallest buildings in their city

- http://en.wikipedia.org/wiki/**Rembrandt**
- Categories: Rembrandt | Dutch painters | Dutch engravers | Dutch Golden Age painters | Baroque painters | Portrait artists | People from Leiden | People from Amsterdam | **Leiden University** | 1669 deaths | 1606 births

W3C Semantic Web for Life Sciences mailing list

Why don't biologists modularise OWL ontologies properly?

Er, well, like how should we do it "properly" and where are the tools to help us?

We don't know and we haven't got any. But here are some vague guidelines.

http://twiki.mygrid.org.uk/twiki/pub/Mygrid/PresentationStore/ISWC2005keynote-final.ppt.

There are no proper ontologies in biology! We have all this incredible stuff in OWL you aren't using. Look at this example I have for you using the Amazon web service

How do I handle my legacy? The web services aren't mine. The ontology is already used

on

Tell them to start again and do it properly this time.

http://twiki.mygrid.org.uk/twiki/pub/Mygrid/PresentationStore/ISWC2005keynote-final.ppt.

Old vs. new

- **Background**: The old issues of knowledge representation and of basic linguistic description are not solved …
  - The bad news: people forget/ignore/despise previous work
- **Innovation**: Follow the new trends and develop applications and programs that were never thought before
  - The good news: a lot of success does not require much knowledge (most users are ignorant anyway)

Evaluation and its value

++ forces one to think about **purposes** and **aims**

-- leads to little creative activity

- 95-97 papers
- A4 format

- commercial/impact value is often due to totally different factors
  - $$$ vs. fame vs. usefulness

Motivation

- *Evaluation is itself a first-class research activity: creation of effective evaluation methods drives more rapid progress and better communication within a research community* (Hirschman, 1998:302f)
- *[Before] there were no common measures and no shared data. As a consequence, systems and approaches could not be precisely compared and results could not be replicated* (Gaizauskas, 1998:249)

Santos 2000, Tutorial om evaluation at SBIA/IBERAMIA

Lack of evaluation history

- In linguistics and the humanities in general
- In computer science
  - *There are plenty of computer science theories that haven't been tested* (Tichy, 1998:33)
  - *I'm not aware of any solid evidence showing that C++ is superior to C with respect to programmer productivity or software quality* (Tichy, 1998:35)
  - *the standard of empirical research in software engineering […] is poor* (Kitchenham et al., 2002:721)

Santos 2000, Tutorial om evaluation at SBIA/IBERAMIA

Food for thought

- The more complex evaluation procedure(s)
  - the more mature is the (sub)area
  - the easier is to progress
  - the easier not to reinvent the wheel
  - the easier to know what one is doing
- Start developing/describing your system by suggesting a way to evaluate it ...

Santos 2000, Tutorial om evaluation at SBIA/IBERAMIA

Anja Belz: "that's nice…"

- In 1981, Sparck-Jones already summarizes 20 years of evaluation in HLT
- We achieved comparability at the price of diversity: the range of evaluation methods has shrunk dramatically
- S&T is littered with the remains of intrinsic measures discarded when extrinsic measures revealed them to be unreliable indicators

Need to evaluate extrinsically

Academia vs. society

Two different models?

- **Scientific excellence**, non-profit, personal value depending on analytical intelligence and scholarly performance
- **Society adequation**, either by state support or by marker regulation, profit-centered, personal value from society role

Now: Model convergence? Or new models emerging?

Funding models

- Bureaucrats decide on what to spend money
  - Top down financing
  - "Objective" value depending on publication weight, impact factors, etc. See The Curious Elightenment of Professor Caritat, Steven Lukes, 1995
- Scientists compete on free funding
  - Bottom-up competition
- Scientists get/use their own funding from other activities

On what a scientific problem is

- The issue of scientific paradigms
- … or research agendas
- There are no objective research paths
  - creativity,
  - methodical doubt,
  - confusion unwrapping
- Hard to distinguish in advance which paths lead to original ideas or impact
  - This is why most project proposals are vitiated

The issue of fashions in science

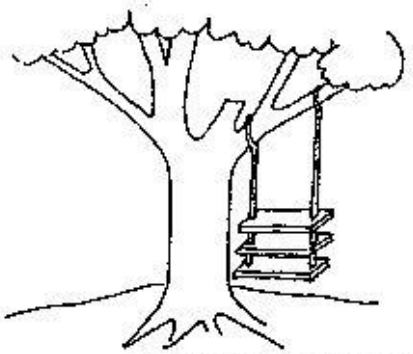
- Mainstream
  - VLSI
  - Web
  - Bioinformatics
- Economic model
  - Antennas
  - Mathematic equation solving for criptography
  - …

The issue of value-added work in a researcher life: some suggestions

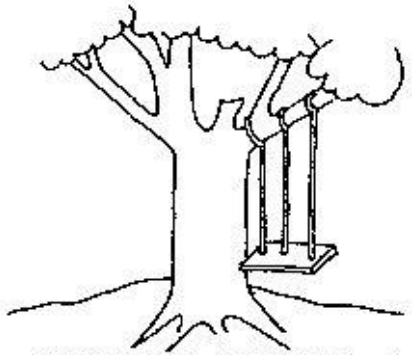
- Journals of negative results
- FAQ /lists and general info on the Web
- Benchmarks and their criticism
- Repeating some studies, checking for hidden assumptions
- Providing the results to the community
- Innovative evaluation procedures and checks
- Log analysis and user studies

Measuring the problem

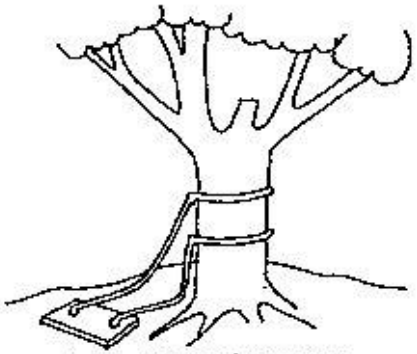
- How often do you hear about **solutions** without a clear understanding and measurement of the problem?
- How often is the "problem" something else?
- How often solving one problem creates another?
- How many empirically sound estimates exist that people can really use, cite, and replicate? Contrary to "opiniated" beliefs…
- What is the (user) tolerance to the problem?



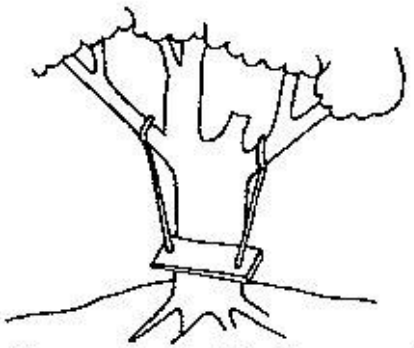
AS PROPOSED BY THE PROJECT SPONSOR



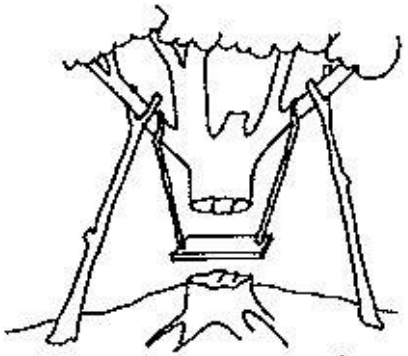
AS SPECIFIED IN THE PROJECT REQUEST



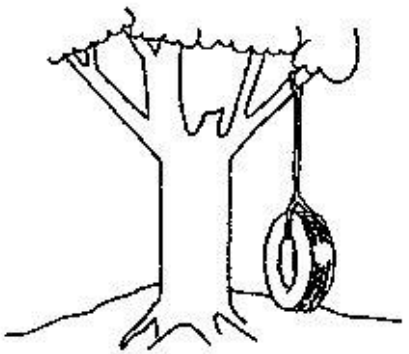
AS DESIGNED BY THE SENIOR SYSTEMS ANALYST



AS PRODUCED BY THE PROGRAMMERS



AS INSTALLED AT THE USER'S SITE



WHAT THE USER WANTED

What the user wanted

A language learner survey by Eric Atwell

- An ELIZA to talk with
- Probably no longer required due to chat rooms and VR
- But ideal for getting material to drain the user and to study his/her weaknesses

The problem of self-image, or field definition

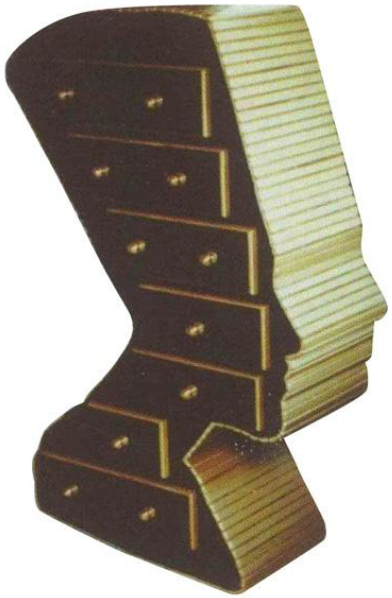
- GIR – GIS – Meteorology – Geology …

… all use maps ☺

- Language and speech
- Medium is essential in the way people are schooled: sound waves, vocal tract morphology, vs cryptography, IR, text processing, vs image recognition

Area displacement

- Almost-random chest of drawers?
- Examples
  - Advanced physics III : stochastic models of population
  - Linguistics for media
- How many times are there new courses based on specific interests of the teacher? (and how many times there are teachers teaching courses which do not interest them?)



http://www.tmjdesigns.co.uk/

Evaluation of publication

Armstrong (1982) studied

objectivity, replicability, importance, competence, intelligibility and efficiency

as criteria for evaluating publication (policy)

Frequently, authors select unimportant problems, advocate a dominant hypothesis, avoid challenges to current beliefs, provide less than full disclosure, use complex methods, or write in a obscure manner (Armstrong, 1982:97ff)

Armstrong (1982)'s findings…

- Referees are seriously biased by knowledge of
  - who wrote the article
  - whether the conclusions of the paper conform with their beliefs (p.85)
- 12 out of 30 replications yielded results that conflicted with the original study (p. 88)
- Journals do not seem to give preference to important problems (p. 90)

As far as refereeing is concerned: my **plea for Signed Reviews**

The politics of publication

What are politicians interested in?

- We are working with RCAAP, maybe we can do this analysis some day?
- Contentwise, it does not matter where you publish
- If you want to have impact, just make a lot of noise in the respective fora, pointing out an obscure paper in an obscure journal
- If you don't want others to have impact, just do not let them publish in the fora you have access, and/or define away the places where it was published

One thing is career advancement, another scientific progress

A true story

- Sebastião J. Formosinho. *Nos bastidores da ciência: resistência dos cientistas à inovação científica*, Gradiva, 1988.
- Different paradigms teach different things, because they see the world in different ways.
- Is the helium atom a molecule? (p.104)
  - Yes (a chemist), No (a physicist)
- *Theories die, not because they were refuted or falsified, but because their proponents and followers die* (Planck)

Concluding appeals

A true researcher:

- Challenges the "scientific" establishment
- Thinks out of the drawer
- Looks at neighbouring disciplines
- Tries to be intelectually honest
- Bewares of fashions and buzz-words

Ellis, John M. *Language, Thought and Logic*. Evanston, IL: Northwestern University Press, 1993.

Q: What is the importance of what you are doing to the big picture?

References

- Armstrong, J. Scott. "Research on Scientific Journals: Implications for Editors and Authors", *Journal of Forecasting* **1** (1982), pp. 83-104.
- Artstein, Ron & Massimo Poesio. "Inter-Coder Agreement for Computational Linguistics". *Computational Linguistics* **34**, 4, 2008, pp. 555-596.
- Atwell, Eric. *The Language Machine: The Impact of Speech and Language Technologies on English Language Teaching*, The British Council, July 1999.
- Belz, Anja. "That's nice ... what can you do with it?", *Computational Linguistics* **35**, 1, 2009, pp. 111-118.

References (cont.)

- Friburger, N. & D. Maurel. “Textual Similarity Based on Proper Names”, in *Proceeedings of Mathematical/Formal Methods in Information Retrieval, SIGIR 2002* (Tampere, 15 August 2002), pp. 155-167.
- Gaizauskas, Robert. "Evaluation in language and speech technology". *Computer Speech and Language*, **12** (4) (1998), pp. 249-62.
- Goble, Carole. “Using the Semantic Web for e-Science: inspiration, incubation, irritation”, keynote at the 4th international Semantic Web conference (Galway, Ireland, 6-10 November 2004), http://twiki.mygrid.org.uk/twiki/pub/Mygrid/PresentationStore/ISWC2005keynote-final.ppt.

References (cont.)

- Hirschman, Lynette. "Language Understanding Evaluations: Lessons Learned from MUC and ATIS", *Proceedings of The First International Conference on Language Resources and Evaluation* (Granada, 28-30 May 1998), pp.117-122.
- Kitchenham, Barbara A., Shari Lawrence Pfleeger, Lesley M. Pickard, Peter W. Jones, David C. Hoaglin, Khaled El Emam & Jarrett Rosenberg. "Preliminary guidelines for empirical research in software engineering". *IEEE Trans. Softw. Eng.* **28**, 8 (Aug. 2002), pp. 721-734.
- Resnik, Philip & Noah A. Smith. "The Web as a parallel corpus". *Computational Linguistics* **29**, 3, September 2003, pp. 349-380.

References (cont.)

- Sjöbergh, Jonas. "Creating a free digital Japanese-Swedish dictionary", *PACLING 2005*.
- Tichy, Walter F. “Should Computer Scientists Experiment More?”, *Computer* **31**, 5, May 1998, pp. 32-40.