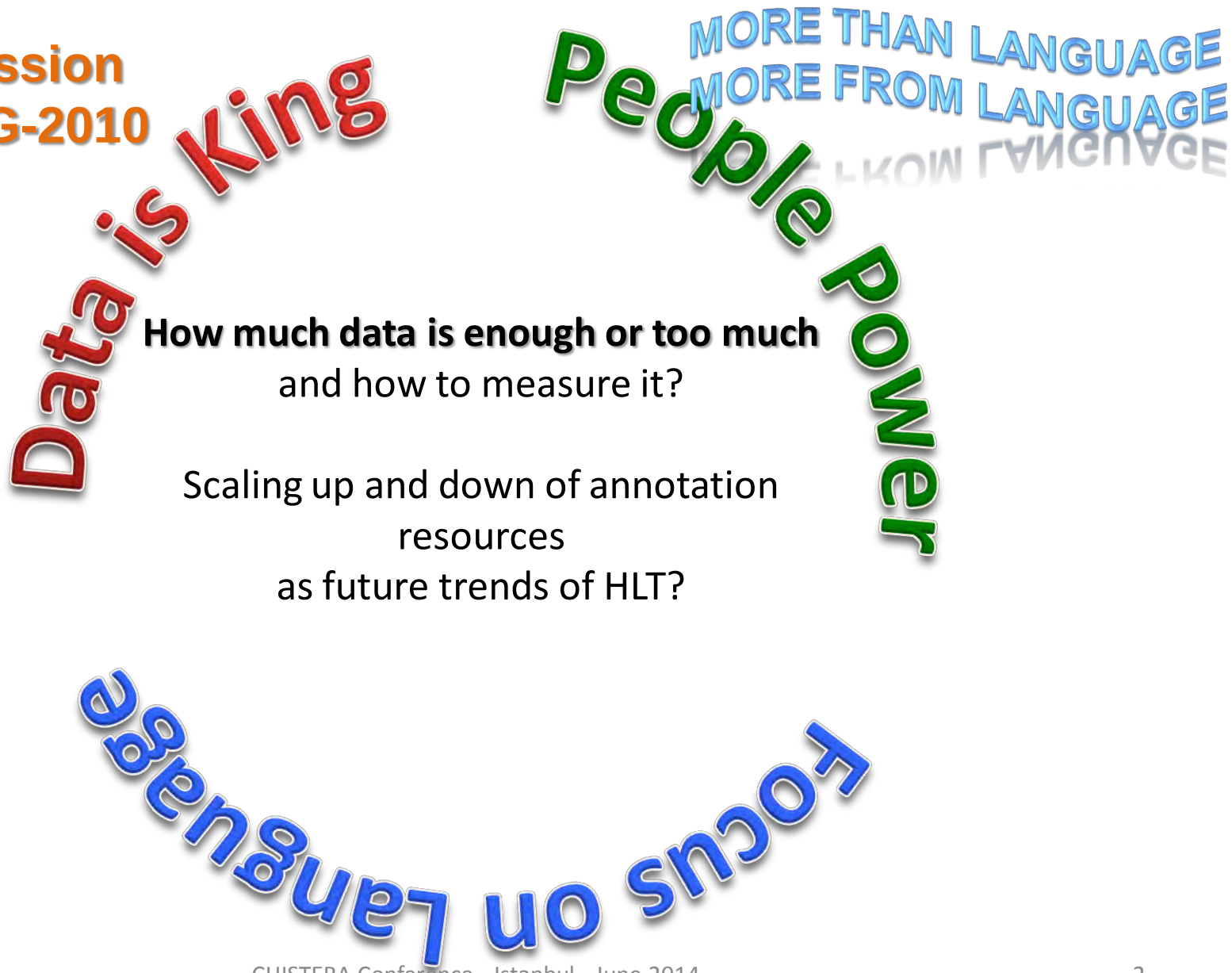
Language Technology Challenges at the Crossroads of Data, Language and Evaluation

Nicoletta Calzolari
ILC – CNR
glottolo@ilc.cnr.it

Nicoletta Calzolari, Chu-Ren Huang, and Dekang Lin

RING Session
@ COLING-2010



How much data is enough or too much
and how to measure it?

Scaling up and down of annotation resources
as future trends of HLT?

NLP is "data intensive"

Story of the last decade is about **Language Resources**

Open Data

Even if in the backstage

Not in the forefront wrt applications

Infrastructural nature

- Important consequences…

Methodological & Policy dimensions

Here in the forefront

Also when dealing with other data types

Old slide with Antonio Zampolli ('80s/early '90s)

Why such needed LRs, are lacking after 30 years of R&D in the field?

→ **1)** Because the main trend until **mid-'80s** was to privilege the processing of ***so-called* "critical" phenomena**, studied by the dominating linguistic theories, rather than focusing on the deep analysis of the **real uses** of a language

- As a result CL was focusing on:
  - ***few examples*** - *often* ***artificially built***
  - *lexicons made of few entries (**toy lexicons**)*
  - *grammars with* ***poor coverage***

→ **2)** Because large-scale LRs are **costly** & their production requires a big organizing effort

Why we still lack them??

… back from the late ’70s/‘80s

Historical notes

Automatic acquisition of lexical information from MRDs

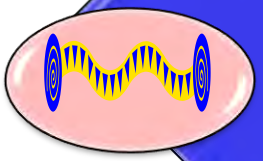
Pioneering Research

- Was my first research & became central in the **Pisa** group (**ACQUILEX**)
- And also **Amsler, Briscoe, Boguraev, Wilks**’ group, **IBM**, then **Japanese** groups, …
- The trend was: “**large-scale computational methods for the transformation of machine readable dictionaries into machine tractable dictionaries**”
- Instead of relying on linguists’ introspection

Evident that:

Part of the results of meaning extraction (e.g. many meaning distinctions generalised over lexicographic definitions and automatically captured)

were ***unmanageable at the formal representation level*** & had to be blurred into unique features and values



Unfortunately, it is **<u>still today</u> difficult to constrain word-meanings within a rigorously defined organization**: by their very nature they tend to evade strict boundaries

… back from the late '80s

After acquisition from MRDs,

Automatic acquisition of info from texts:

This trend has become today a consolidated & pervasive fact

- From acquisition of **"linguistic information"**
- To acquisition of **"general knowledge",** with more data intensive, robust, reliable methods

Lesson learned

Going from core sets **to large coverage** has implications not just in quantitative terms, but more interestingly in terms of **changes to the models and the strategies of processes**

Lesson learned

Need of adequate models to handle actual usage of language

We started building:

LRs as necessary "infrastructure"

both for research & applications

LRs give to NLP systems
the **knowledge needed** for the various linguistic processing

Realising that most of the needed information

- **escapes individual "introspection"**
- can only be **acquired** analysing large **textual corpora attesting language use** in different fields/communicative contexts

- **Sub-product?:** Importance of **statistical** methods

Preamble

- We wanted more & more data …

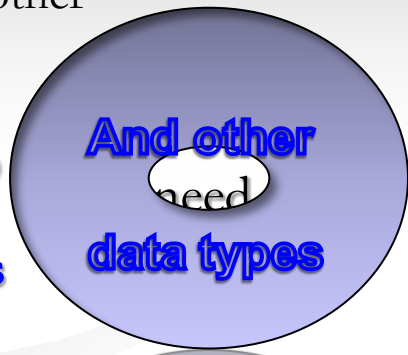
Have we been too successful ?!?

- We experience today a sort of statistical "intoxication" !
  - It started as a new strategy, a revolution maybe? But it has turned to tactics. Stuck with it? In a narrow loop of small advances, not linked to each other
- Can we add also a new strategy? and hopefully a vision?

Main Statement

- We tend to forget about "language"

  understand its properties & complexities



- Where do we (try to) encode what we know about language properties?
  - In annotations

BUT

- *Is there any theory & methodology of studying and exploiting*
- *the interactions among the various annotation layers?*

Vision

- *Like the big Genome project, … a large Language initiative*

What we need

If we have the purpose of modelling human language communication

- Some careful thinking about an overall computational theory of semantics, anchored on **language communication**
  - How we might create **large corpora** containing it … i.e. **beyond textual corpora**
- How the different NLP applications might use them & on them our methodologies
- Share methodologies for LR creation: **Portability across modalities**: of models, technologies, methodologies, approaches, ...
- How improve our overall understanding of language

BIG DATA Also Visual, Multimodal

LR Methodologies
for other Data types
in contexts of real life

Keywords:

- **LR sharing/linking/integrating/reusing/ benchmarking …**
- **Content interoperability** → towards **Knowledge Resources**
- Paradigm of **accumulation of knowledge,** so successful in more mature disciplines, allowing sharing of multi-level processed/annotated resources
  - Putting in place new ways of collaboration, also **among communities**
  - **Principle of** processing → **re-depositing**

Collaborative building of LRs

- **A Unified Framework for LRs & other data types**
- **Adopting consolidated methodologies**

Evaluation

Interoperability

Infrastructural issues

Language resources for action modelling in robotics

Language resources (such as WordNet, VerbNet, etc.) and ontologies **contain world/commonsense knowledge** that can be used for action modelling in robotics:

- for robot-human interfaces that enable communication in realistic settings
- to **extract knowledge** about the **world** and use it in reasoning modules
- to **bridge the gap between the semantic content of images from computer vision** and **what we know about objects** typically involved in actions

- High standardisation **reusing language foundations**, tools, applications
- Incremental design & development

FLaReNet Recommendations: a global perspective

International Cooperation

INFRASTRUCTURE

Sustainability

Recognition

Development

Documentation

Interoperability

Availability

Coverage/Quality

DATA

Resource Coverage, Quality, Adequacy

"Address appropriate coverage in terms of quantity, quality and adequacy to technological purposes"

- **Facts**
  - In current **data-driven paradigm, innovation** depends on **big amounts of data**, of the **right type,** appropriate **quality,** & for **more languages**
  - **Careful:** Dependence on data may create **disparity** for **under-resourced languages** and domains
- **Actions to be taken**
  - **Increase quantity of resources** available to address language & application needs
  - Provide **high-quality resources for all European languages**
  - Implement **BLaRKs** for all languages, especially less-resourced languages
  - **Basic and long term research** on **automatic production of LRs**
  - More efficient uses of annotated data with **repurposing & merging** techniques

DATA

Resource Coverage, Quality, Adequacy

"Address appropriate coverage in terms of quantity, quality and adequacy to technological purposes"

- **Actions to be taken (cont)**
  - **Inter-disciplinary research** with other scientific domains working on large volumes of data
  - Address **formal and content quality of resources** by promoting **evaluation** and **validation**
  - A **European evaluation and validation body:** an infrastructure for coordinated LRT evaluation & validation
  - Establish **common and standard Language Technology evaluation procedures**
  - Define and establish a **Quality Seal of Approval,** to be endorsed by the community

Interoperability Conditions

Technical

- ***Common Metadata***
- ***Explicit semantics*** *of metadata*
- ***Semantic interoperability***
- ***Tools*** *that help using standards*
- ***Operational*** *standards*

Infrastructural

- *A* ***common repository of standards***
- *Common* ***documentation templates***
- *A framework that facilitates testing*
- *An* ***interoperability framework for/of web services – standards as services***

Social/Cultural

- ***Community involvement***
- *Disseminating but not forcing*
- *Interoperability as academic research area*
- ***Link to sustainability***

Resource Availability

"Make resources easily and readily available, within an adequate IPR and legal framework"

- **Facts**
  - **Reluctance** in fully embracing an open data model is still common
  - Intellectual Property Rights (**IPR**) issues are crucial to facilitating growth in our sector
  - **Legislation** is lagging behind the technology
- **Actions to be taken**
  - Opt for **openness** of LRs, especially **publicly funded** ones
  - **Clear IPR** at the **early stages** of production; try to ensure that re-use is permitted
  - Elaborate specific, **simple** and **harmonised licensing solutions** for data resources

USE

Services

For a European service platform

- LRs as services
  - Composite access
  - Web-services for Visualisation, Analysis, ...
  - Extracting, Adapting, Merging, Linking, ...
- Services around LRs
  - Sharing: Authentication, ...
  - Legal
  - Web-services for Collecting, Crawling, Cleaning, Linking, Integrating, Clustering, ...
  - Inventorying
  - Describing with MD
  - Converting (around Interoperability)
  - Annotating , (Content) Analysing, Acquiring info, ...
  - Adapting , Repurposing, Evaluating,
  - Crowdsourcing
  - Translating, Localising, ...
  - Summarising , Mining, ...
  - Understanding, ...
- For whom?

From language neutral ...

Towards more language specific (basic and advanced)

Towards more language-based

NEW

Topics: towards real life!

- **Multimodal/Multimedia data** (video, image)
- **Space : Spatial** info is very complex, is pervasive & involves many levels of info. Links with **Time, Subjectivity, ...**
  - Capture the complex constructions of spatial language: corpus based
  - Inventory of how spatial information is presented in language
  - Lend itself to Mechanical Turks
  - Many **Multimedia applications**
- **New social media**
- **Social Intelligence** – Technologies for e-participation & decisional support
- **Subjectivity, emotions, ...**
- **Crowdsourcing, Games with a Purpose, ... Human Computation!!** (crowd inside processing)
  - Links to Social & Behavioural sciences, Interactive systems, Usability, Ethics, ...

NEW

Improve semantics??

- **Learn concepts/ontologies from new types of (large) corpora**
  - To improve the semantic capabilities in "usual" NLP tasks
  - Improve learning from other semantic/pragmatic contexts
  - Mix different modalities
  - Multimodal ontologies, with new properties …
- **Get data from new types** (also sensors, GPS, ..)
  - ... And **see if semantic models improve**, if able to fill the gaps/errors unsolvable when remaining in the same paradigm
- **Big Tasks goals tested in classical evaluation**
  - With evaluation methodologies appropriate for the task
- **Different communities involved** (social aspect)

Vision: LT as a « mature science »

This vision has to do with many LR-related dimensions even more important in a **data-intensive** discipline like LT

- Shifting to a culture of ***sharing***, ***re-use***, ***reproducibility of research results***
- Changing the way of ***making science***, in a sort of light revolution towards openness of science, to become a **mature science**
  - The small cost that each of us will pay to document, share, etc. should be paid back benefiting of others' efforts

- Leading to a greater ***opportunity*** of ***collaboration***, encouraging ***bigger experiments*** by larger collaborative teams (as more mature sciences)
- Moreover, reproducibility encourages **trust**

BIG MORE

Recognise the Value – & solve the Conflict

No Strategic Plan can do without

Look also at other disciplines

Processed Data is power

Infra struc tural

Compe tition

Public concerns

Data policies

Best way to capture the full potential

Economic benefits

BIG USE MORE

The Challenge: how to unlock the value