Uno sguardo al passato

... pensando al futuro

Policy Issues

surrounding Language Technology:

why should we care?

Relevance & challenges of meta-research issues:

openness, interoperability, collaboration, integrity, reproducibility, ...

Nicoletta Calzolari

ILC – CNR

glottolo@ilc.cnr.it
The beginnings: From the end of the ‘40s to the beginning of the ‘60s

- The uses of “computational techniques” for processing linguistic data can be distinguished in two main directions:
  - The attempts of **Machine Translation (MT)**
    - 1948: MT at MIT & National Physical Laboratory, see Booth 1958
  - **Electronic processing of texts**, supporting research in the Humanities
    - 1948: beginning of the *Index Thomisticus*, by R. Busa

---

... But soon **Divergence between two areas**
Historical Overview

1948-51 First experiments of automatic text processing on Latin texts (St. Thomas - Father Busa)

1958-60 First application of automatic processing of Italian = lexical analysis and phonological statistics (thesis by A. Zampolli)

1965 Inauguration of CNUCE - PISA = Concordances of Divine Comedy (IBM)

1968 All the Italian projects of the field (linguistics, philology, lexicology, lexicography, philosophy, etc.) are in Pisa = Zampolli (IBM) founds the Linguistic Division of CNUCE

1968 First proposal of an electronic lexical resource (Zampolli at the Conference “Lexicon Electronicum Latinum - Pisa)
Historical Overview – cont.

1970-72  LIF (Frequency Italian Lexicon)
        First International Conference of the field in Italy (Linguistics and Computers)
1970    First official University course in Computational Linguistics (Pisa)
1972    Creation of the DMI (Italian Machine Dictionary), for an Information Retrieval project of the Chamber of Deputies
1973    COLING in Italy (Pisa)
1978    The CNUCE Linguistic Division becomes the Institute of Computational Linguistics of CNR
1981    European Science Foundation Conference, with representatives of all the European Research Funding Agencies = first proposal of international standards for linguistic data
New developments

From the mid ‘80s two new paradigms start and soon become established:

- **Language Industry** (Vidal-Beneyto 1991)
- **Reusable Language Resources** (Grosseto 1986) (Walker, Zampolli, Calzolari, 1995)

This favours a renewed **collaboration between the two areas, HTP & NLP**

- To exploit the **complementarity** of their methods and knowledge
  - corpora, style/genres, lexicons, etc.
- Thus encouraging the possibility to find new types of **funding** from government agencies in all parts of the world
1986 Workshop “On Automatic the Lexicon” (Grosseto) Zampolli, Walker & Calzolari

→ a new paradigm (in Kuhn’s sense) is introduced in the international context of CL

A Manifesto was produced where

→ central role of Language Resources in CL (infrastructural, polytheoretical, multifunctional, reusable) was emphasised

→ a number of actions were recommended: the foundations for a large number of initiatives which took place later in Europe

→ the data-driven approach is brought to new life, besides the “rule-based” one (imposed by the “Chomskian effect”)
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

- After many years of disregard – even disdain & contempt – for LRs, due mainly to the prevalence and influence of the generativist school

Automatic acquisition of lexical information from MRDs

- Was a new idea & my first research & became central in the Pisa group (ACQUILEX) – with much less powerful tools than now
- Also Amsler and then Briscoe, Boguraev, Wilks, IBM, Nagao, …
- The new trend was: “large-scale computational methods for the transformation of machine readable dictionaries into machine tractable dictionaries”
ACQUILEX: Extraction from texts vs. formal representation in lexicons

- The rigour & lack of flexibility of a TFS (typed feature structures) representation language can cause difficulties when mapping into it NL words, ambiguous and flexible by their own nature.

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 still today difficult to constrain word-meanings within a rigorously defined organization: by their very nature they tend to evade strict boundaries.

N. Calzolari 9 Euralex - 1996
After that pioneering era, production & use of adequate LRs strongly increased

The **lexicon** has become ever more relevant

- Both international and national authorities started investing in the field as never before, interested in technologies & systems which are really working & are economically interesting

- The need of **empirical methods**, based on the analysis of large amount of data, has been recognised

- LRs must be **robust enough** for analysing the concrete uses of a language, either theoretically “interesting” or not
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

(IN-)Adequacy of (current) lexicons
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
All started with the situation we had in the late ‘80s – early ‘90s

With all the **XXX-LEX projects**

- GeneLex
- MultiLex
- AcquiLex

A. Zampolli: *Let’s be coherent!*

**EAGLES**

**ISLE**

**Standards, Best Practices, ...**