Divergences betw. Lexicon encoding & Corpus annotation

In the **lexicon**
- senses are “de-contextualized” (a necessity to capture generalizations)
- sense discrimination must be kept “under control” → clustering (manually or automatically)

In the **corpus sense annotation** task
- contextualization plays a predominant role
- calls for a range of **pragmatic issues**
- corpus analysis per se would lead to excessive **granularity of sense distinctions**

- Capture just the core basic distinctions in a core lexicon &
- Acquire additional, more granular info (usu. of **collocational** nature) from corpora

  to be encoded within the broader senses, e.g. to help translation
A number of “dichotomies” not as opposite views, but as complementary perspectives

Language as a *continuum*:

- rules vs. tendencies
- absolute constraints vs. preferences
- discreteness vs. continuity/gradedness
- theoretical/potential vs. actual
- intuition/introspection vs. empirical evidence
- theory-driven vs. data-driven
- symbolic vs. statistical

Choices on the *syntagmatic axis* are pervasive

Lexicon & Corpus must converge
Correlations between different levels of linguistic description in the design of a Lexical Entry

To understand word-meaning:

✔ focus on the correlation between syntactic & semantic aspects
✔ but other linguistic levels - morphology, morphosyntax, lexical cooccurrence, collocational data, etc. - are closely interrelated/involved
✔ these correlations must be captured when accounting for meaning discrimination
✔ the complexity of these interrelationships makes semantic disambiguation such a hard task in NLP

← textual corpora as a device to discover and reveal the intricacy of these relationships
← Frame/SIMPLE semantics as a device to unravel and disentangle the complex situation into elementary and computationally manageable pieces
Mismatch between LRs and LT

- Often a gap between advancement in LRs and LT
- Either adequate LRs are missing … or there are no systems able to use “knowledge intensive” LRs effectively
- Shortcomings:
  - lack of usable implementations fully exploiting new types of LRs
  - LR claims are not empirically evaluated

A parallel evolution of R&D for both LRs and LT is needed

This is not yet happening today, and requires ➔ more overall coordination
Lack of communication betw. the communities of HLT & Semantic Web/Ontologists

The Semantic Web (SW) needs HLT
&
HLT will highly benefit from the SW

…otherwise risk of re-discovery of what was done 20 years ago…

see first issue of the *International Journal on Semantic Web & Information Systems, 2005,*

with statements identical to ours in papers of the ‘80s!!!
In the Semantic Web vision ...

...need to tackle the twofold challenge of

- content availability &
- multilinguality

Natural convergence with HLT:

- multilingual semantic processing
- ontologies
- semantic-syntactic computational lexicons
Issues in LR & LT research agenda converging with SW needs

From LT:
- Meaning & content → Knowledge
  - Semantic markup, personalised annotation, semantic content retrieval
- Discover relationships
- Semantic lexicons/ Terminologies/ Ontologies
  - Integration, mapping, modelling, …

Viceversa, from SW:
- Collaborative & distributed infrastructure
- Interoperability & standards
  - Extending current extraction/annotation capabilities:
    - to other media
    - to dynamic processes
As a critical step for **semantic mark-up** in the Semantic Web
Infrastructure of Language Resources...

...static

♦ Semantic networks (e.g. Euro-WordNet)
♦ Lexicons & Ontologies
♦ TreeBanks (with many layers of annotation)

...dynamic

♦ Lexical acquisition systems from corpora
♦ Infrastructure of tools
  ♦ Robust morphosyntactic & syntactic analysers
  ♦ Word-sense disambiguation systems
  ♦ Sense classifiers
  ♦ ...

But... they will never be “complete”
The Infrastructural Role of LRs Is Widely Recognized

➔ to sketch the *situation in Europe* wrt implementing this infrastructure, underlining conditions specific of the European approach

The message I want to convey:

➔ there has been an *underlying global strategy* behind the set of initiatives launched in the last years in EU
➔ a *global vision is necessary* to achieve more coherent and useful results
➔ *outstanding aspects* that need *further common effort*

Heidelberg, July 2000
History: ILC & international LRs initiatives

- ACQUILEX [since ’88]
- MULTILEX
- ET-7
- ET-10
- TEI
- NERC
- RELATOR
- ONOMASTICA
- MULTEXT
- COLSIT
- LSGRAM
- DELIS
- EAGLES
- PAROLE
- SIMPLE
- SPARKLE
- ELSNET
- EuroWordNet
- MATE
- NITE
- Cluster 488 (Italian)
- TAL (Italian)
- ISLE
- ENABLER
- INTERA
- LIRICS
- …
- Senseval/Semeval
- WRITE
- Forum TAL (Italian)
- …
- ISO
- ELRA
- LREC
- LRE Journal
- NEDO
- Language Grid
- BootStrep
- KYOTO
- …

Essential role of EC to start a basic Infrastructure

Established a model

EU at the forefront in the areas of LRs and standards in the ’90s
Today: a broad “potential” Infrastructure
Vitality & Success signs... for LRs

RELATOR
EAGLES/ISLE
ENABLER
ELSNET
TELRI
INTERA
LIRICS
...
ELRA
BLARK
Unified Lexicon (W/S)

LREC
LRE journal
...
ERANET-LangNet
...

EU

Internat

LDC & others
ISO
COCOSDA/_WRITE
US Cyberinfrastructure
Japan COE21
NEDO
Language Grid

Cooperative initiatives – Links to...

FLaReNet (ICT)
CLARIN (ESFRI)

National
Moreover …
Many dimensions around the notion of language

Political issues
- e.g. a commonly agreed list of minimal requirements for “national” LRs: BLARK

Cultural issues
- Language and cultural identity
- Language and humanities

Need of bodies for a broad research agenda & strategic actions for LT&LRs (W/S/MM)

Technical issues
- Multilingualism
- Economic, social issues
  - Application
  - Services