Do conditions exist for standardisation work?

- **Reusability** as key concept:
  - to avoid duplication of efforts, costs, etc.
  - to allow synergies, integration, exchange of data

- The feasibility of formulation of consensual standards as a **strong sign of maturity** in the field

- EAGLES was launched in ‘93
Which Standards?

➡ De facto standards, based on consensus, pooling results from current efforts/projects

➡ Implications on work methodology (bottom up) and on organisation of work (widely distributed, with cooperation between industry and academia)

EAGLES is acting as a catalyst to help the LE community crystallise descriptions of problem areas.
Harmonised Language Resources

- PAROLE Corpus
- PAROLE Lexicon (morphology and syntax)
- SIMPLE Lexicon (semantic subcategorisation)
- EuroWordNet Lexicon (semantic network)

- Extended in National Projects

- annotated corpus, grammars, annotated dialogues,
- used in applications

- True Infrastructure of Language Resources
- Basis for Multilingual Resources

N. Calzolari
Prague, March 2002
First attempt to tackle encoding of semantic subcategorisation frames on a large scale, for so many languages

The multidimensional characterisation of senses provides multiple ways to capture locally salient common features

also having challenging research aspects

and providing a framework for testing and evaluating the maturity of the current state-of-the-art in lexical semantics

Potential basis for future European multilingual initiatives for NLP and LE applications
Harmonisation: Need for a Global View

✔ The various components (data & software) interact with each other

✔ Need for compatibility

Formalisms
Grammars
Software: Taggers, Chunkers, Parsers...

Languages

Representation
Lexicon

Software: Acquisition
Systems
I/O Interfaces

Annotation
Corpus

NC Heidelberg, July 2000
Vital, for a sound LR strategy, to accompany core static LRs with dynamic means for enriching & integrating them - on the fly - with the types of information structurally and intrinsically missing. This global view eliminates an apparent dichotomy: static vs. dynamically built (or incremental) resources in a more comprehensive perspective that sees the two as complementary.

Lexical semantics extracted from text is the challenge from structure to contents.

Heidelberg, July 2000
Do Computational Lexicons faithfully represent the reality of language in use?

- or - as it is often the case with traditional dictionaries - they represent a sort of stereotypical/theoretical language?

Carefully constructed/selected large corpora are essential sources of linguistic knowledge

- for the extensive description of the concrete use of the language in real text,

impossible relying on introspection only and on native speakers’ (even if linguists/lexicographers) intuition

evidence of actual usage is frequently in contrast with what one would expect based only on introspection
Lexicon and Corpus: a multi-faceted interaction

- L → C: tagging
- C → L: frequencies (of different linguistic “objects”)
- C → L: proper nouns, acronyms, ...
- L → C: parsing, chunking, ...
- C → L: training of parsers
- C → L: lexicon updating
- C → L: “collocational” data (MWE, idioms, gram. patterns ...)
- C → L: “nuances” of meanings & semantic clustering
- C → L: acquisition of lexical (syntactic/semantic) knowledge
- L → C: semantic tagging/word-sense disambiguation (e.g. in Senseval)
- C → L: more semantic information on LE
- C → L: corpus based computational lexicography
- C → L: validation of lexical models
- C → L: ...
- L → C: ...

Prague, March 2002
One of the most interesting - and intriguing - aspects of corpus use for a lexicographic task: the **impossibility** of descriptions based on a **clear-cut boundary** between what is **admitted** and what is **not**

In actual usage, language displays a large number of properties behaving as a **continuum**, and not as properties of "yes/no" type

The same is true for the so-called “rules”, where we find more of a “**tendency**” towards a **rule** than a precise rule in corpus evidence

Most of the lexical (grammatical, syntactic, semantic) information is **not constraining**, but rather **preferential**. (This creates problems at the level of the formal representation, which must be able to accommodate this preferential information: this may not be easy and certainly not straightforward for a constraint-based formalism)
Consequence of the **corpus-based approach**

- **Break hypotheses** too easily taken for granted in mainstream linguistics
- **Many properties** behave as a continuum, not as “yes/no” properties
- So-called “rules” are simplifications or idealisations dispelled by **real usage**, more frequently “tendencies” towards a rule

---

**Lesson learned**

`[IN-]Adequacy of Lexical resources`

A long way to be able to **recognise & integrate all the dimensions** relevant to **content** interpretation
Constraints or Preferences?

• Problem of the distinction between constraining and preferential information:
  • not inherent in the nature of the data, but
  • related to their use: the same specifications (e.g. number or voice) can be used either as constraints or as preferences
    • Preferential information is connected more to specific attributes:
      – usage indices specifying the register, style, variant;
      – type of action expressed by the verb;
      – sortal semantic restrictions on complements and adjuncts, and on adjective collocates

A lot of information should be treated
• not as absolute constraints, whose violation makes a sentence totally unacceptable,
• but as preferences, making a given sentence more or less acceptable in a given context without affecting its grammaticality
A computational lexicon has to represent “irregular” facts: Induction of knowledge from texts

- **divergences of actual usage** from what is **potentially/in theory acceptable**

- represent - and distinguish - what is **allowed, but only very rarely instantiated**, with respect to what is both **allowed and actually used**

  ➔ more robust and flexible tools are needed for **(semi)automatic induction of linguistic knowledge from texts**

  ➔ some bootstrapping methodology

**Lexical semantics extracted from texts is the challenge ➔ from structure to contents**

Baslow, 1999
towards Corpus based Semantic Lexicons … at least in principle

- both in the design of the model, &
- in the building of the lexicon (at least partially)
  - with (semi-)automatic means

Design of the **lexical entry** with a combined approach:

- **theoretical**: e.g. Fillmore frame semantics/generative lexicon, ...
- **empirical**: Corpus evidence

✔ **even if**: not always there are sound and explicit criteria for classification according to “frame elements”/qualia relations/…