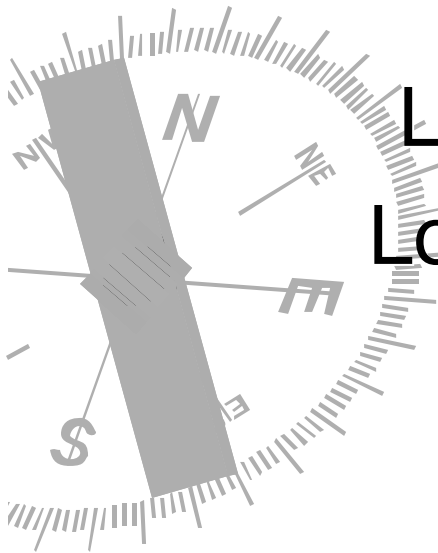


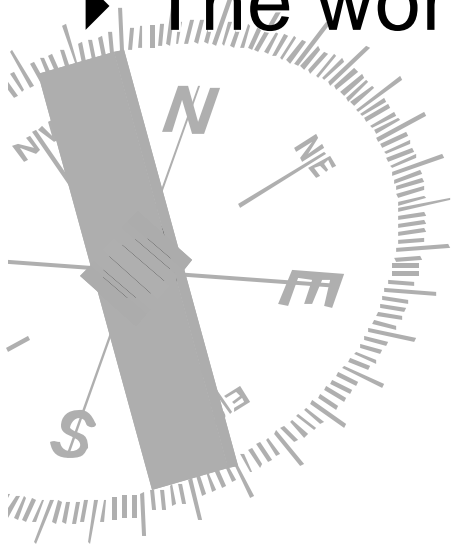
Towards a roadmap for standardization in language technology

Laurent Romary & Nancy Ide
Loria-INRIA — Vassar College



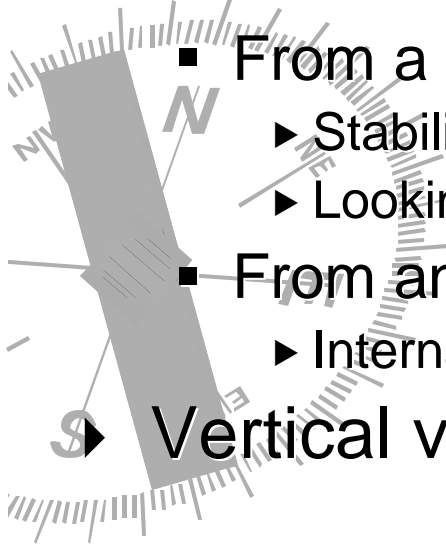
Overview

- ▶ General background on standardization
- ▶ Available standards
- ▶ On-going activities
- ▶ The work ahead of us



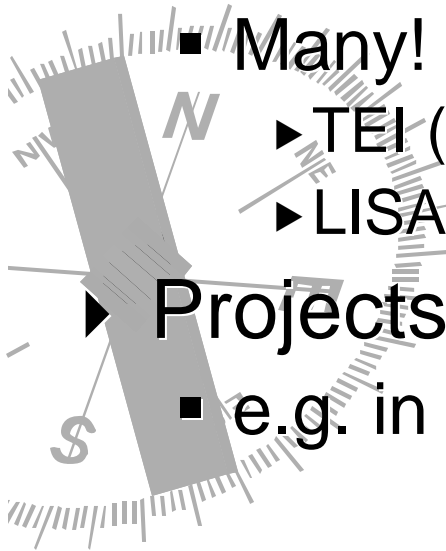
Standardization

- ▶ Defining methods or models to facilitate
 - Exchange of data
 - Interoperability between software components
 - Comparability of results
- ▶ Involves
 - From a technological point of view
 - ▶ Stabilizing existing practices
 - ▶ Looking ahead for potential roadblocks
 - From an organizational point of view
 - ▶ International consensus, long term availability and maintenance
- ▶ Vertical vs. horizontal standardization



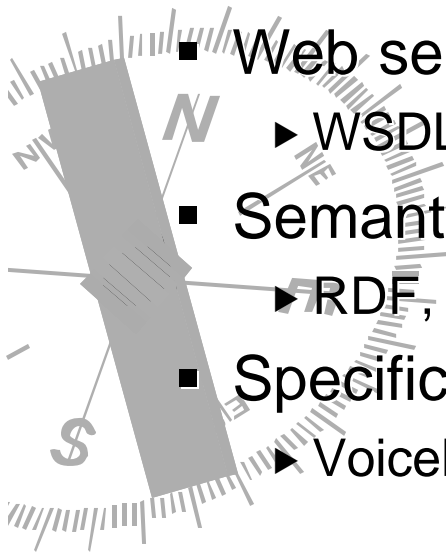
Standards: a complex picture

- ▶ Official standardization bodies:
 - National: AFNOR, ANSI, DIN, BSI, MSA
 - International: ISO, IEC, CEN, W3C, OASIS
- ▶ Specific fora:
 - Many! e.g.:
 - ▶ TEI (Text Encoding Initiative)
 - ▶ LISA (Localization Industry Standards Association)
- ▶ Projects with a pre-normative purpose:
 - e.g. in EU: EAGLES, Multext, MATE, ISLE



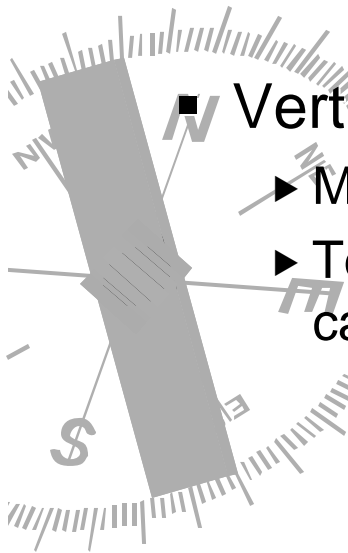
Existing standards (1)

- ▶ W3C (World Wide Web consortium); horizontal standards
 - Basic building blocks:
 - ▶ XML, XML Schemas (Note: growing importance of alternative RelaxNG schemas), XSL
 - Web services activity
 - ▶ WSDL, SOAP
 - Semantic web activity
 - ▶ RDF, RDFS, OWL
 - Specific (vertical) activities with little critical mass
 - ▶ VoiceML, EMMA, etc.



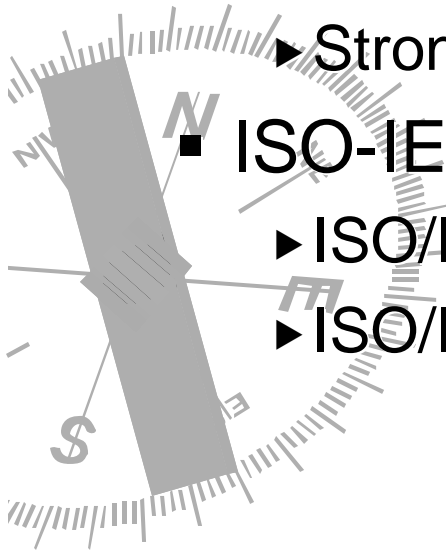
Existing standards (2)

- ▶ Relevant standards in ISO (partial view)
 - Basic infrastructural (horizontal) standards
 - ▶ Character encoding (cf. IPA): ISO 10646/Unicode
 - ▶ Language codes: ISO 639 (e.g. 'fr') and ISO 639-2 (e.g. 'fra'/'fre')
 - Note: under ISO/TC 37/SC 2
 - Vertical standards
 - ▶ MPEG7 for multimedia information — hardly implementable :-(
 - ▶ Terminology standards: ISO 12200 (Martif), ISO 12620 (Data categories), ISO 16642 (Terminological markup framework)
 - Note: under ISO/TC 37/SC 3



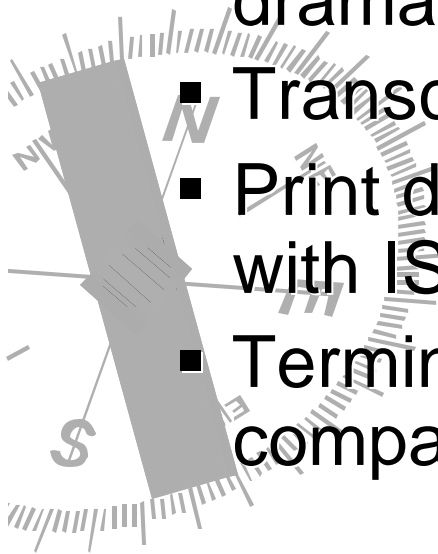
Existing standards (3)

- ▶ Looking at other fields
 - ISO-IEC/JTC 1/SC 36: education
 - ▶ Collaboration on language aspects
 - ISO-IEC/JTC 1/SC 32: databases
 - ▶ Strong basis provided by ISO 11179
 - ISO-IEC/JTC 1/SC ??: evaluation of software
 - ▶ ISO/IEC 9126-1 [2 & 3 in progress]
 - ▶ ISO/IEC 14598-1 to 6



Existing standards (4)

- ▶ TEI proposals relevant for our field:
 - TEI header: seminal work to evolve in collaboration with IMDI and OLAC
 - Basic representation of texts: prose, poetry, drama, etc.
 - Transcription of speech
 - Print dictionaries: under revision in collaboration with ISO/TC 37/SC 4 (cf. LMF)
 - Terminologies: under revision to make it compatible with ISO 16642



ISO committee on language resources

- ▶ **ISO TC37** - Terminology and other language resources

- **SC3** - Computer applications in terminology

- ▶ ISO 12200 - Martif

- Latest version of TEI Terminology chapter

- ▶ ISO 12620 - Data categories (under revision)

- ▶ ISO 16642 - TMF (Terminological Markup Framework)

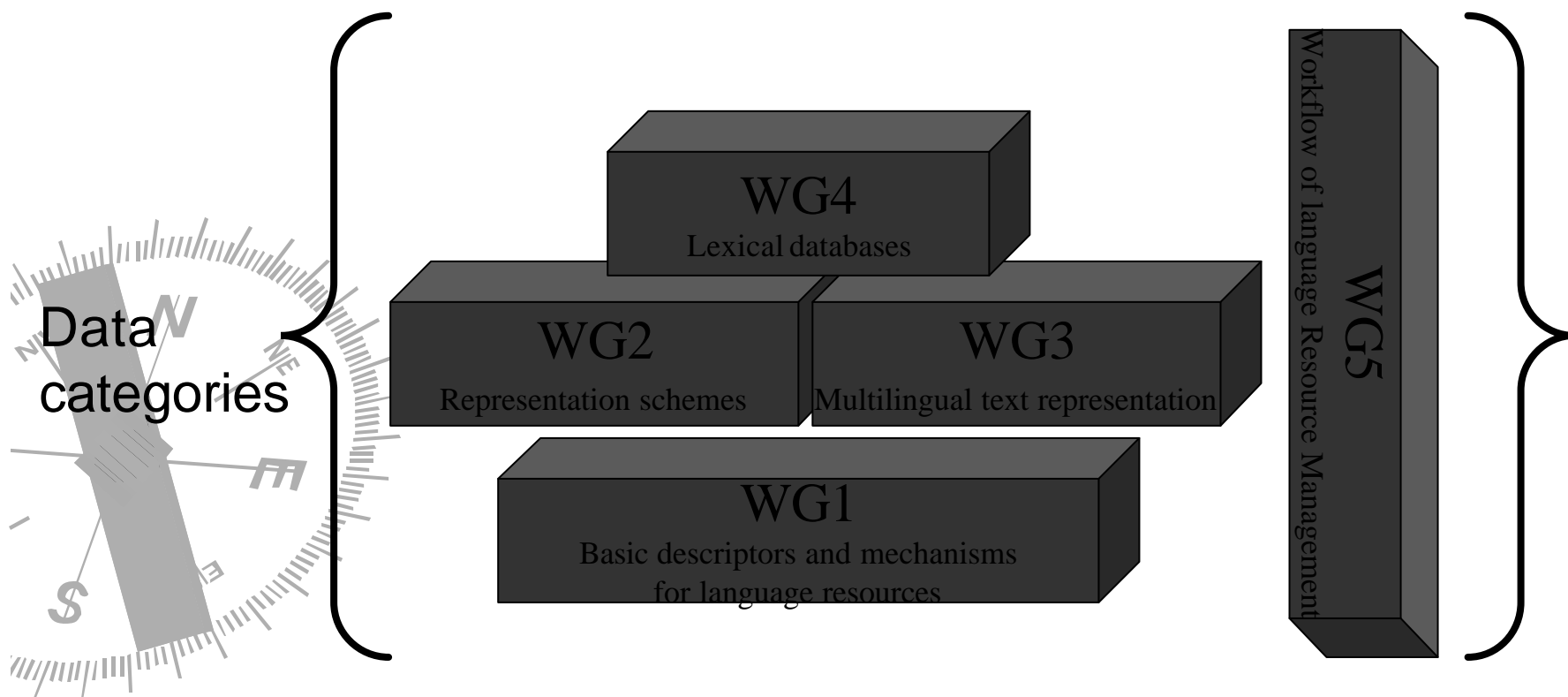
- **SC4** - Language Resource Management (May 2002)

- ▶ Sec.: K.-S. Choi, Chair.: L. Romary

- ▶ <http://www.tc37sc4.org>



ISO/TC 37/SC 4 overall rationale

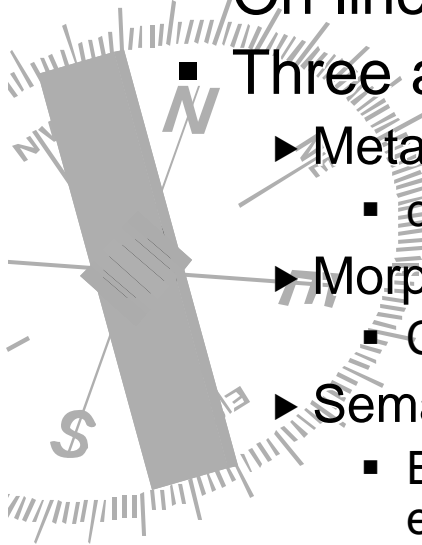


On-going activities within ISO/TC 37/SC 4 (1)

- ▶ Feature structure representation
 - Joint activity with the TEI; CD document almost achieved; planned project on FS declaration
- ▶ Linguistic Annotation Framework
 - E.g. principles of annotation scheme specification and representation, pointing mechanisms for stand-off markup; draft document available
- ▶ Morphosyntactic annotation framework
 - Stable working draft under dissemination for evaluation
- ▶ Lexical Markup Framework (LMF)
 - A general specification platform for lexical structures
 - Preliminary proposals: core model + lexical extensions

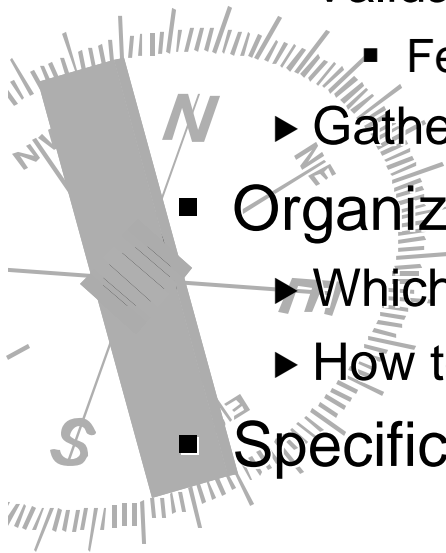
On-going activities within ISO/TC 37/SC 4 (2)

- ▶ The central role of the Data Category Registry
 - Objective: market place of descriptors for all types of language resources and annotation schemes
 - ▶ E.g.: /grammatical gender/, /paucal number/, /ablative case/, etc.
 - On-line tool available: <http://syntax.loria.fr>
 - Three ad hoc groups created
 - ▶ Metadata for language resources
 - cf. TEI, IMDI, OLAC
 - ▶ Morphosyntactic descriptors (SC4 plenary last Tuesday)
 - Cf. Morphosyntactic Annotation Framework
 - ▶ Semantic content descriptors
 - Exploratory: discourse relations, dialogue acts, referential links, etc.



Priorities for the future (1)

- ▶ Stabilizing and disseminating
 - Wide dissemination of existing standards
 - Two priorities in ISO/TC 37/SC 4: morphosyntax and lexical structures
 - ▶ Validation of on-going documents by our community
 - Feedback on documents, reference implementations
 - ▶ Gathering up samples and/or test suites (manpower needed)
 - Organizing the work on the Data Category Registry
 - ▶ Which additional topics should be addressed?
 - ▶ How to involve a wide variety of experts?
 - Specific publication and information days



Priorities for the future (2)

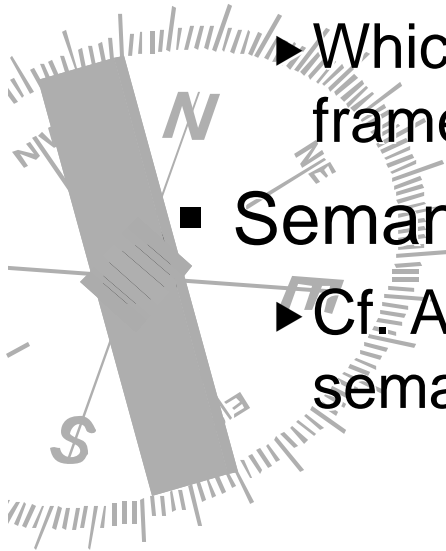
▶ Filling in the gaps:

- Syntactic structures: cf. Treebanks, (Chunk, deep) Parsers
- Application specific lexica

▶ Which formats should be 'frozen' within the LMF framework

▪ Semantic content representation

▶ Cf. ACL/SIGSEM working group on Multimodal semantic content representation



Priorities for the future (3)

▶ Open fields

- Multilingual information representation

- ▶ How to relate on-going activities on translation memories, localization, iTV, multimedia information (e.g. sub-titling)

- Evaluation of NLP components

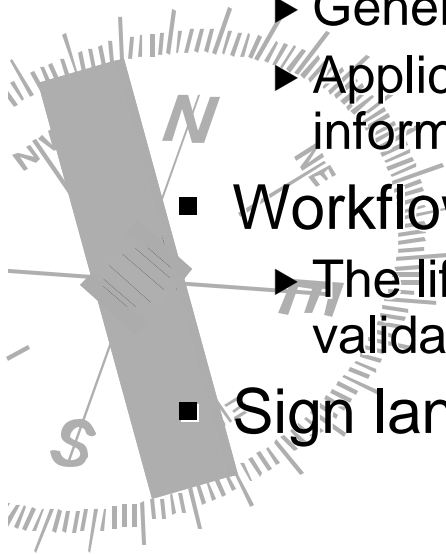
- ▶ General principles: linguistic coverage, metrics

- ▶ Application specific evaluation methods: machine translation, information extraction

- Workflow of language resources

- ▶ The life cycle of language resources: creation, enrichment, validation, dissemination

- Sign languages...



Conclusion

- ▶ Importance of dissemination of existing standards (in academia...)
 - Standards as the identification of stable concepts in a field
 - Introduction in academic curricula
- ▶ Importance of wide involvement of experts (academia and industry)
 - Defining priorities
 - Contribution to technical work
- ▶ Linking main milestones in the roadmap with the underlying standardization efforts
 - E.g. Evaluation related standards