LR&E
The perspective of an H(L)T developer

Fabio Pianesi

ITC-irst

pianesi@itc.it

What LR and evaluation do we rely upon?

- Mostly collected or built on site (but also resort to anything is available and can be useful).
- Driven by the technologies we are working at (hence, needs can change accordingly)
- Current major interests are in

  Automatic WSD,

  Terminology Extraction,

  Multilingual (STST) and Multisensorial Communication,

  Information Presentation

Lexical resources

- MultiWordNet – same synsets for Italian and English.
  - Version 1.37
    - Word Senses: 58731
    - Words: 41989
    - Synsets: 32536
- List of 4.000 English-Italian lexical gaps.
- List of 20.000 Italian collocations
- Form-based lexicon for Italian (approx. 500.000 forms)

Grammatical Resources

- In use
  - Small Functional Systemic Grammar for Italian. For NLG purposes (Ilex/Exprimo)
  - Large Unification Grammar for Italian (analysis)
- Planned
  - Italian and French FUF grammars for the museum domain
  - Corpus focusing on spatial expressions (locations, directives)
    - Spoken corpus, with transcription
    - Interactions between two subjects
    - Scenario: museum and cultural heritage.

Written corpora

Si-TAL Treebank

- 220,000 words general balanced corpus, and 90,000 words specialized corpus for the financial domain
- Annotations: orthographic, morphosyntactic, syntactic (constituency and functional), and semantic (word senses).

Written corpora …. continued

- Corpus “YY"
  - from year 1992 to year1999
  - Approx. 200.000.000 tokens
- Corpus “XX"
  - From year 1985 to year 2001
  - Approx. 300.000.000 tokens
- Corpus “ZZ“. Alligned, bilingual (Italian/English) news
  - Form October 2000.
  - Approx. 2.000.000 + 2.000.000 tokens

Written corpora ….continued

- C-STAR - Alligned multilingual (Japanese, English, Italian and Germa) corpus
  - Phrase books sentences.
  - English as pivot language.
  - Japanese: approx. 200.000 sentences, Italian 27.000 available (full translation under way).
  - Use: statistical machine translation

Written corpora – future plans

- MEANING (FP5) corpus, consisting of
  - 42 balanced domain-specific corpora,
  - 1.000.000 tokens each
- plus
  - a general balanced corpus (Italian).
  - La Repubblica, La Stampa, Vita Trentina (contact under way)
  - 100.000.000 planned tokens
- Multilevel annotation: structure of the texts (i.e. the primary data), orthographic features, morphosyntactic information, collocations, named entities, and word senses.

Spoken corpora

- IBNC (ELRA, LRP&P) - Radio broadcast news
  - Duration: total 31h:13m, speech 29h:57m (all transcribed)
  - Use: automatic transcription
- Video broadcast
  - Duration: 66h:12m (transcribed: 32h:45m)
  - We continue to record news
  - Use: automatic transcription
- Speech Recognition for data entry (SPEEDATA)
  - Duration: 5h:45m (all transcribed)

Spoken corpora ….continued

- APASCI – fonetically balanced utterances (WoZ)
  - Duration: ~ 3h:30m (all transcribed)
  - 16,090 utterances and digits.
  - Use: ASR training
- DIGIT
  - Duration: ~10 hours isolated and connected digits (all transcribed)
  - Use: ASR training
- PHONE1-PHONE2 (telephonic speech)
  - Duration: 7 hours with phonetic rich sentences (all transcribed)
  - Use: ASR training, dialogue over telephone

Spoken corpora ….continued

- FIELD - telephonic speech
  - Duration: ~ 80 hours (alphadigits, names, ...). Partly hand-transcribed, partly automatically transcribed.
  - Use: ASR training, dialogue over telephone

- CARINI – fairy tales and short stories
  - Duration: ~ 1 hour speech (all transcribed)
  - Transcription + syntactic (constituency) annotation
  - ToBI, POS, and syntactic (constituency) labeling ongoing
  - Use: prosodic synthesis for TTS and CTS.

Spoken corpora ….continued

- Children read speech L1 and L2 corpus (under construction)
  - Languages: English, German, Italian
  - Use: ASR training for L1 and L2

Multilingual Dialogues

- C-STAR! – Tourism information
  - Duration: ~ 29 h. (all transcribed)
  - Languages: Italian, German, English, French.
  - Use: Speech to speech translation
- NESPOLE! – Tourism information
  - English (37), French (31), German (62), Italian (61)
  - All transcribed and annotated (IF)
  - Use: Speech to speech translation

Multilingual Dialogues

- NESPOLE! Multimodal and multilingual dialogue corpus
  - Collected with real system
  - Duration: 16.5 h
  - English, German and Italian
  - Pen-based gestures
- NESPOLE! – Expanded Tourism Domain. Both H323 and clean speech
  - English (16), French (16), German (16), Italian (16)
  - Includes gestures with pen or mouse on maps
  - All transcribed, annotation (IF) almost completed.
  - Use: speech to speech translation

Minimal LR set for each language

- Spoken corpora for ASR training
- Treebanks
- Multilevel (POS, syntax, topic/focus, TOBI, emotions, …) annotated corpora
- Comparable / aligned multilingual corpora
- Lexica (WordNets?)
- Resources for NLG: lexica, grammars, corpora for canned texts and/or macronodes (shallow approaches to) generation, rhetorical relations, ….. both for development and evaluation

Minimal LR set for each language ….continued

- Multilevel annotated video-audio corpora: emotions, gestures, posture, other elements of human behaviour (complex sequences).

Multicultural issues

- Corpora from different target populations (children, the elderly, people with special needs)

Current programs

- Multilingual (STST) and multi-modal/-sensorial dialogues
  - Emotion and prosody (annotated spoken corpora for the extraction and synthesis of appropriate cues)
  - Emotions and facial/visual cues (annotated recordings, movies, talk shows, ….).
  - Language and other behavioural cues: gestures, posture.
  - Multilingual (and multimodal) corpora targeting children

Current programs…. continued

- Automatic extraction of lexical information
  - General WordNets
  - Domain lexica/ WordNets
  - Terminology extraction
  - Alligned corpora
- User-centred assessment of technologies /applicative scenarios
  - Usability and cognitive impact/load assessment

Impediments and challenges

- Financial.
- Legal - copyright issues for written corpora; privacy issues for spoken (and other kinds of behavioural) corpora.

Impediments and challenges …continued

- Technical
  - Availability of off-the-shelf tools for transcription/ annotation (POS, constituency, automatic transcribers, …)
  - Annotation standards
  - Effective and agreed upon evaluation procedures
  - LR and evaluation methodologies for NLG

Impediments and challenges …continued

- Emerging issues – short term
  - Multilingual and multisensorial communication, (e.g., address the issue of communicative effectiveness in Multilingual Dialogue)
- Emerging issues – medium/long term
  - Cognitive and behavioural issues/data (and their evaluation, issues relating to experimental design, ..)
    - Can cognitive science contribute to shape/choose among technologies and applicative scenarios?
    - Can we think of behavioural/cognitive data as new information to rely upon while developing technologies/ scenarios?
    - Do we need (multistratal?) behavioural corpora distinguished for languages and/or cultures?
  - new populations (children, the elderly, people with special needs)

Impediments and challenges …continued

- People
  - Strong multidisciplinarity:
    - linguistics,
    - computer sciences,
    - cognitive and behavioural sciences,
    - experiment design.
  - training

Success in international sharing/cooperation

- Most of our activities on LR have been carried on through cooperation within national or international consortia (notice, in most cases the consortium's main goal was not LR production):
  - C-STAR
  - NESPOLE! (FP5/NSF)
  - PF-STAR (FP5)
  - MEANING (FP5)
  - PEACH (PAT)
  - VICO (FP5)
  - IBNC (ELRA, LRP&P)
  - TAL (MUIR)
  - M-PIRO (FP5)
  - WebFaq (PAT)

Success in international sharing/cooperation

- And with companies
  - MultiWordNet (FST)
  - Video and radio broadcast news (RAI)
  - Speech in car (FIAT, BOSCH)
- The role of national and international agencies
  - Add competition to cooperation.
  - (Stricter) cooperation between national and international agencies.
  - Keep up with (and foresee) emerging issues.

Future plans - conclusions

- Provide technological baselines for core technologies:
  - Evaluation: agreed procedures, scenarios and data
  - Competitive assessment (of data and technologies)
  - FP6
- Broaden the target population
  - Children
  - The elderly
  - People with special needs (sensorial/motor/cognitive impairments, special needs for special situations)

Conclusions

- User-centred concerns: human behaviour, cognitive science.
- Which models for resources production/ distribution:
  - centralised,
  - based on agencies (e.g., ELRA)
  - open source-like: impact on standards; trade-offs between collaboration and competition.