a very brief history of the field

1. 1945–1965:
   ▶ machine translation
   ▶ analysis of literary and historical texts
2. 1965–1985:
   ▶ symbolic NLP
   ▶ corpus linguistics
3. 1985–: lexical resources
4. 1995–: data-driven NLP
5. 2005–:
   ▶ WWW-driven NLP
   ▶ ontology resources
   ▶ serious standardization efforts

key issues in LTR building

▶ LTR building should be based on commonly accepted standards
▶ build a core set of LTR, designed in a harmonized way for all languages (the “BLARK”)
▶ formulate a distribution policy in order to make LTR available to the community

the big picture

▶ Int. conference on language resources and evaluation (LREC; most recently in Istanbul 2012)
▶ Language Resources and Evaluation journal (Springer)
▶ numerous satellite conferences/workshops at COLING, ACL, etc. (see the ACL Anthology)
▶ distribution agencies: Linguistic Data Consortium (LDC; USA); European Language Resources Association (ELRA; Europe)
▶ LT infrastructure networks: CLARIN; FlarReNet; SILT; . . .
▶ standardization: ISO TC37/SC4
LTR according to CLARIN

- annotations
- lexical resources
- written corpora
- speech and multimodal resources
- tools

resource building

- the prototypical and central resource is the corpus (in a wide sense)
- this is because the corpus is the closest thing we have to primary language data, which other resources ultimately must refer to and rely on
- corpus resource building is, basically, a process of refinement:
  1. collect text/speech
  2. annotate (somehow) with level-1 annotations
  3. use the text/speech plus the level-1 annotations to annotate the text/speech with level-2 annotations
  4. etc.
  5. etc.

resource annotation

- the crucial part is the annotation (anybody can collect lots of text, if there is text to be had at all)
- annotation requires (linguistic) knowledge
- annotation types are (largely) language-independent
- annotation content is language-dependent
where does the linguistic knowledge come from?

▶ there are basically two modes for acquiring the requisite linguistic knowledge:
  1. manually (writing grammars, etc.)
  2. automatically from annotated (supervised machine learning) or unannotated (unsupervised machine learning) data
▶ (of course, mixed modes are also possible)
▶ in both cases, resources come from resources:
  “For whosoever hath, to him shall be given, and he shall have more abundance: but whosoever hath not, from him shall be taken away even that he hath.” (Matt. 13:12)

manual linguistic knowledge acquisition

▶ requires experts (linguists) who are familiar with the conceptual apparatus and formalism used
▶ uniformly labor-intensive for real-world applications (as opposed to “toy”/proof-of-concept ones), but the effort can be amortized if consistent attention is paid to interoperability and distribution issues
▶ (potentially) better coverage of marginal – linguistically interesting – cases than automatic approaches
▶ exact but brittle
▶ tends to produce “human-type” errors

supervised automatic acquisition

▶ requires experts (linguists) for annotating enough – generally large amounts of – training data, who must be familiar with the annotation scheme (and its conceptual apparatus)
▶ one-off labor-intensive for real-world applications; the effort can be amortized if consistent attention is paid to interoperability, representativeness and distribution issues
▶ (potentially) better coverage of central – linguistically unexciting – cases than automatic approaches
▶ robust but inexact, although for some tasks – e.g., POS tagging – outperforms manual approaches
▶ tends to produce “funny” errors

unsupervised automatic acquisition

▶ in theory no involvement of human expertise; the machine learning algorithm is supposed to infer the classes and their distribution from raw data
▶ requires large amounts of data
▶ the pure unsupervised mode tends to be much less accurate than supervised machine learning
▶ otherwise like supervised machine learning
evaluation and LTR

- In all cases, evaluation is considered de rigueur (although it is logically somewhat contradictory to require it in the unsupervised scenario).
- Evaluation requires a gold standard, a manually annotated or manually corrected representative data set.
- For some types of annotation, “shared task” events are organized (e.g., the CoNLL series of conferences).
- Representativity is an issue which deserves more attention (see, e.g., Giesbrecht/Evert).

evaluation methodology

- Evaluation methodology in the LTR field is in need of development:
  - Commonly, convenience gold standards are used, which more often than not have been developed for other purposes than evaluation.
  - Gold standards are convenience samples with respect to genre, modality, “annotation ontology”, etc.; e.g., a newstext corpus syntactically annotated using binary-branching phrase-structure trees.
  - There has been very little systematic investigation of the influence of, e.g., genre, text type, language, etc., on the accuracy of annotation tools.
  - Given that annotations are based on other annotations, errors may be compounded in an annotation pipeline; again, this issue deserves more study.

corpus resources and tools

“So, a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration.” (McEnery and Wilson 2001, Corpus linguistics: 32)

corpus resources and tools (2)

Text corpus ‘dimensions’:
- Purpose (and reusability)
- Language(s), language variety
- Sampling, representativeness
- Size
- Annotation (“interpretation”)
- Ethical and IPR issues
corpus resources and tools (3)

PURPOSE:
▶ NLP
▶ machine learning
▶ evaluation of NLP systems
▶ linguistics
▶ empirical investigations of language
▶ applied linguistics
▶ investigations of learner language
▶ language learning and testing
▶ language planning and maintenance

corpus resources and tools (4)

empirical investigations of language:
(Literary and Linguistic Computing • Corpora • International Journal of Corpus Linguistics / ALLC • ACH • ICAME)
▶ lexicon (monolingual, translations)
▶ grammar (through function items and grammatical annotation)
▶ sociolinguistics
▶ text linguistics, style, genre
▶ historical linguistics
▶ language typology

corpus resources and tools (5)

investigations of learner language:
▶ learner corpus investigations (lexicon, grammar): over- and underuse
▶ contrastive corpus linguistics: comparing learners’ IL (interlanguage) with their L1 (native language) and with natively produced L2 (the target language)
⇒ make (better) learner models possible in intelligent CALL (Computer-Assisted Language Learning) systems

language learning and testing:
▶ corpus-based exercises (cloze, word class assignment, syntactic analysis)

corpus resources and tools (6)

language planning and maintenance:
▶ basis for lexical resources
▶ basis for LT applications
▶ basis for CALL applications
▶ status-raising
Corpus resources and tools (7)

Language(s), language variety:
- modality (writing, speech, gestures)
- genre, text type
- monolingual, bilingual, multilingual

Corpus resources and tools (8)

Sampling, representativeness:
- a little bit of everything (Brown [Frown], LOB [Flob], SUC, etc.)
- one text type (newspaper text, technical text, etc.)

Size:
- fixed (but bigger all the time!)
- growing (“monitor corpus” [Sinclair]; Wikipedia; Korp’s blog mix)
- dynamic/virtual (“web as corpus”)

Corpus resources and tools (9)

Annotation (“interpretation”):
- (corpus/text metadata)
- transcription (of speech, gestures)
- tokenization and normalization
- part of speech tagging
- lemmatization
- syntactic analysis
- coreference
- alignment of parallel texts
- (search/browsing/concordancing using the annotations)
- etc.

Corpus resources and tools (10)

Some ethical and IPR (Intellectual Property Rights) issues:
- who owns the language? (the appropriation issue)
- are IPR universal? (codified law vs. tradition)
- who benefits from the corpus?
- research ethics
- anonymization (text/audio/video)
- under what terms can we use/publish the corpus? (cf. BNC/Wortschatz/Korp/Google n-grams)