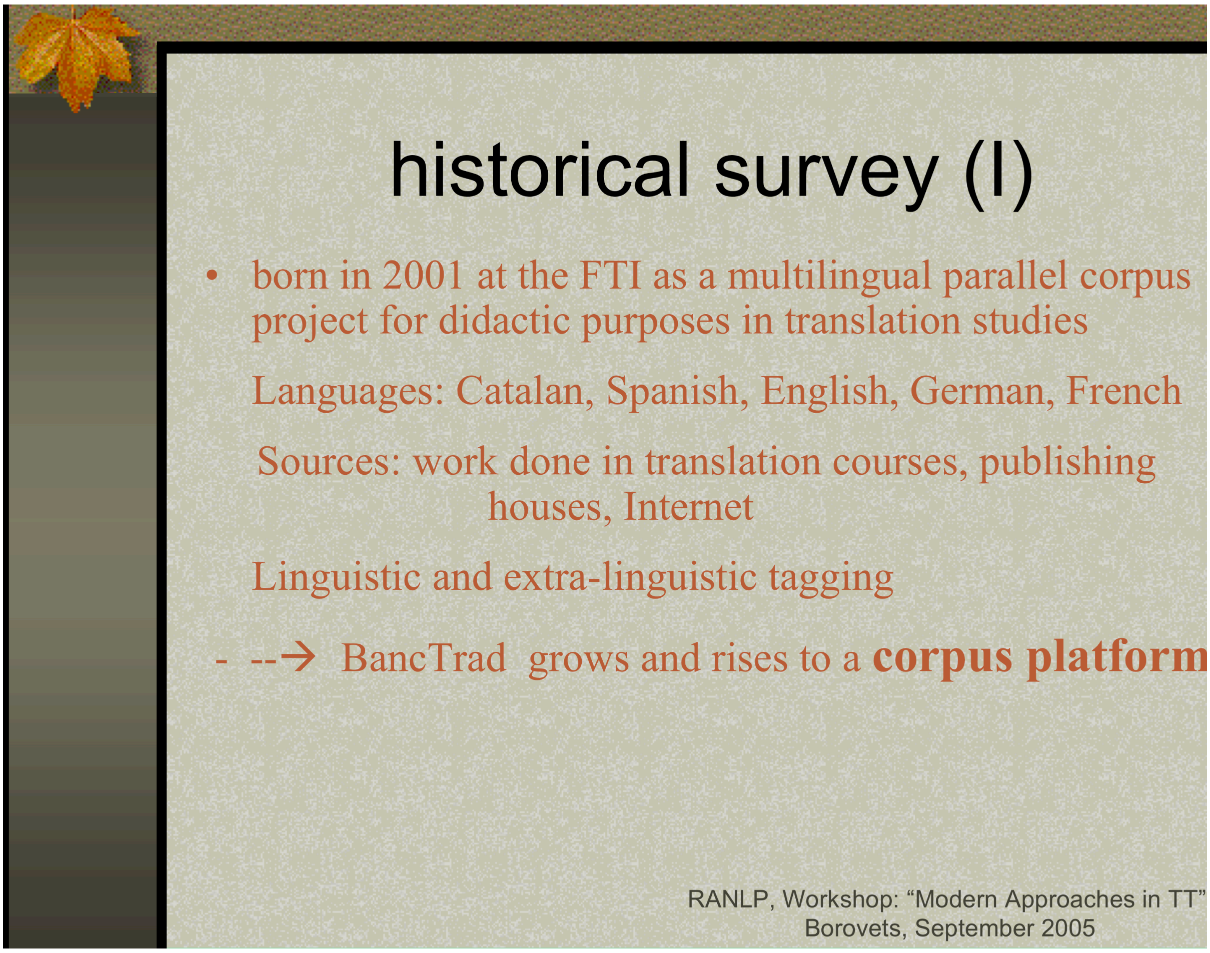
BancTrad: a web interface fo integrated access to annotated corpora

Carme Colominas (carme.colominas@upf.edu)
GliCom (Grup de Lingüística Computacional)
Departament de Traducció i Filologia
Universitat Pompeu Fabra

contents

- description of the BancTrad platform
  - general data
  - linguistic processing
  - architecture
- demo of the input interface
- demo of the query interface



historical survey (I)

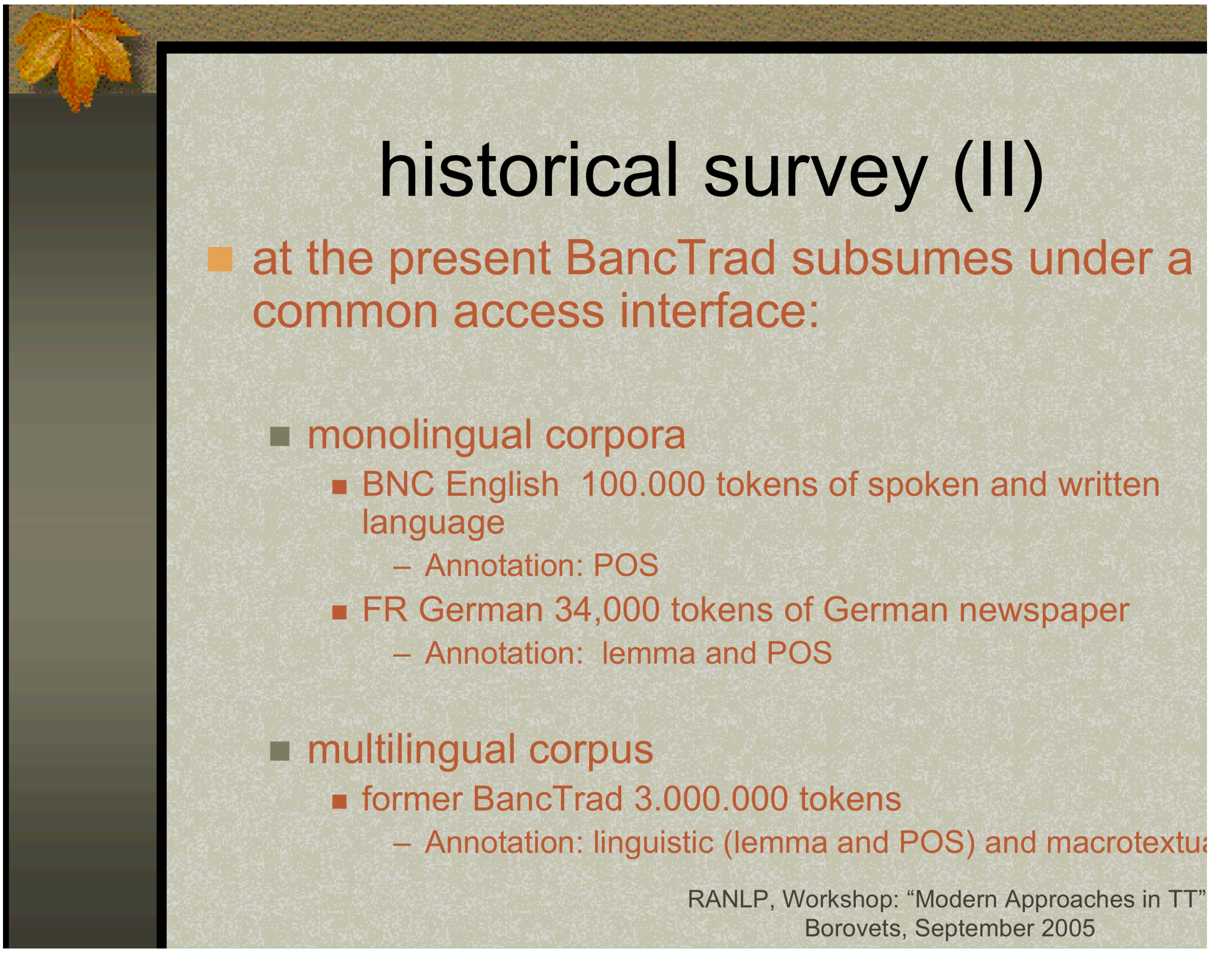
- born in 2001 at the FTI as a multilingual parallel corpus project for didactic purposes in translation studies

Languages: Catalan, Spanish, English, German, French

Sources: work done in translation courses, publishing houses, Internet

Linguistic and extra-linguistic tagging

- --→ BancTrad grows and rises to a **corpus platform**



historical survey (II)

- at the present BancTrad subsumes under a common access interface:
  - monolingual corpora
    - BNC English 100.000 tokens of spoken and written language
      - Annotation: POS
    - FR German 34,000 tokens of German newspaper
      - Annotation: lemma and POS
  - multilingual corpus
    - former BancTrad 3.000.000 tokens
      - Annotation: linguistic (lemma and POS) and macrotextua

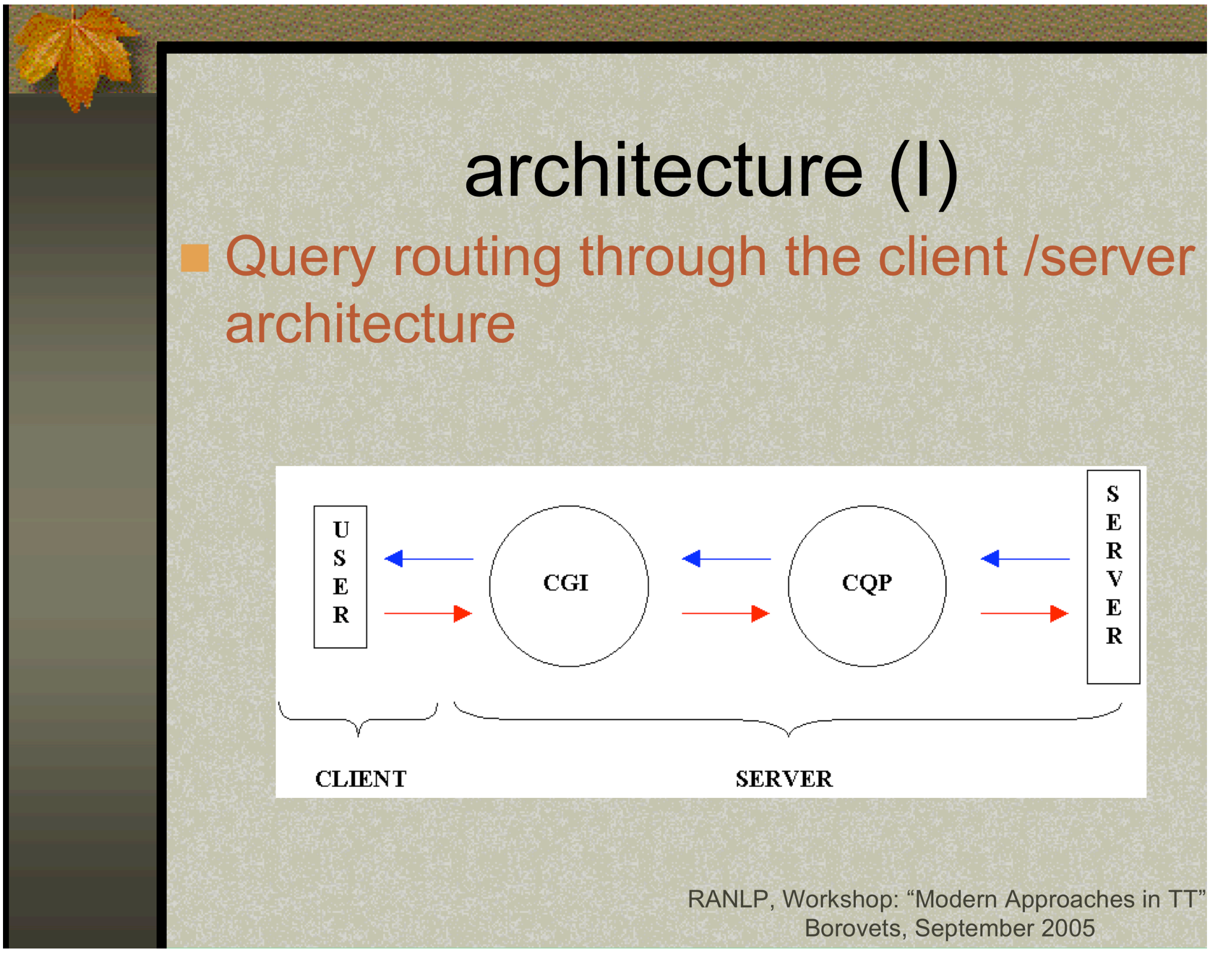
linguistic and extralinguistic processing (I)

- preprocessing via web (Java Web Start application)
- macrotextual mark-up:
  - source and target languages
  - register
  - type of text
  - ...
- output: SGML marked document
- conversion to plain text, alignment and uploading

linguistic and extralinguistic processing (II)

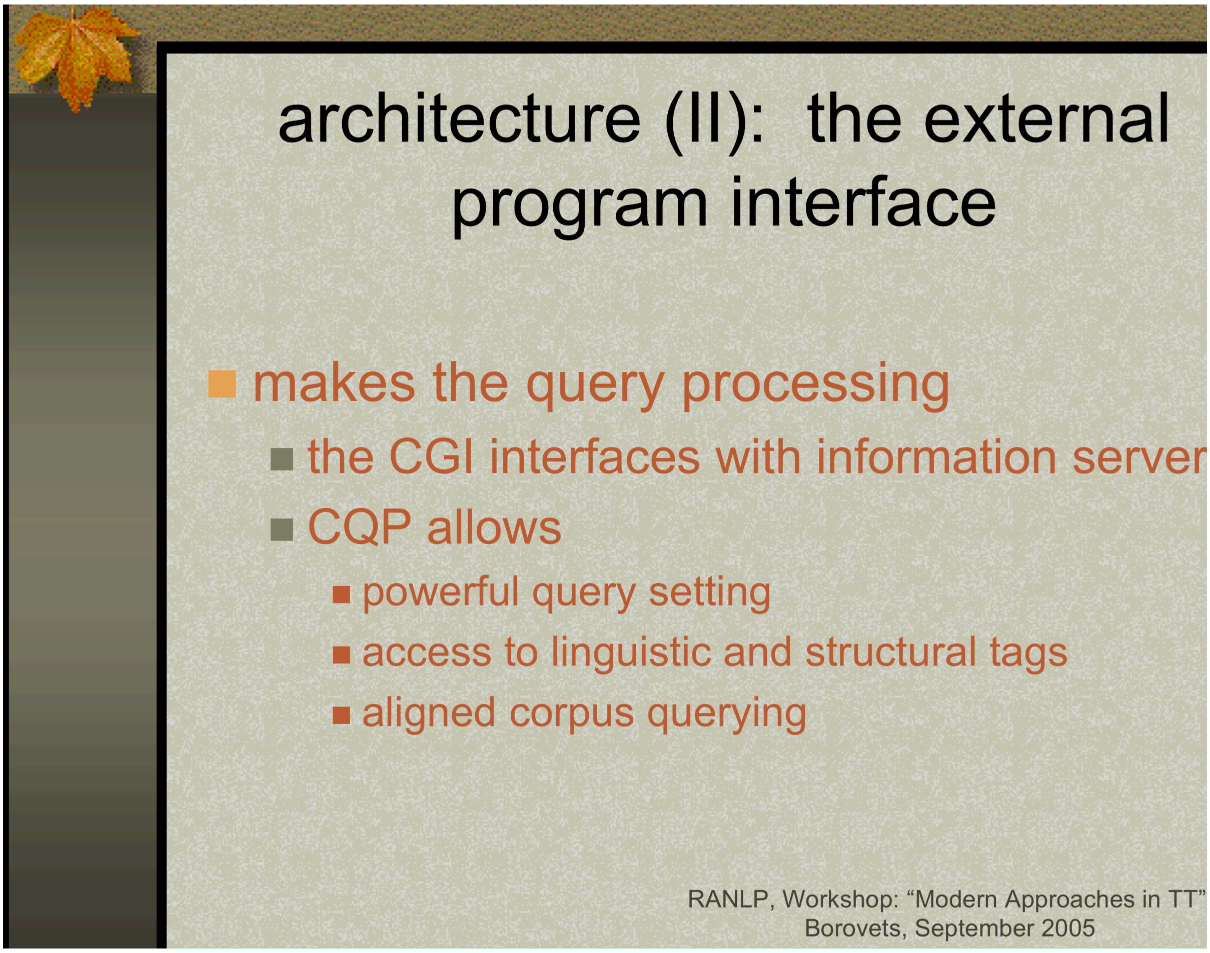
- Tagging:
  - Spanish, English, German, French TreeTagger (Schmid 1995,19
  - Catalan CATCG (Badia *et al*.. 2000)
  - minimum of uniformed kind of information (lemma, POS)
- Corpus formatting with the CWB tools (IMS, Stuttgart)
- Corpus building

  : ready to be consulted with CQP through the GUI and the CGI



architecture (I)

- Query routing through the client /server architecture



architecture (II): the external program interface

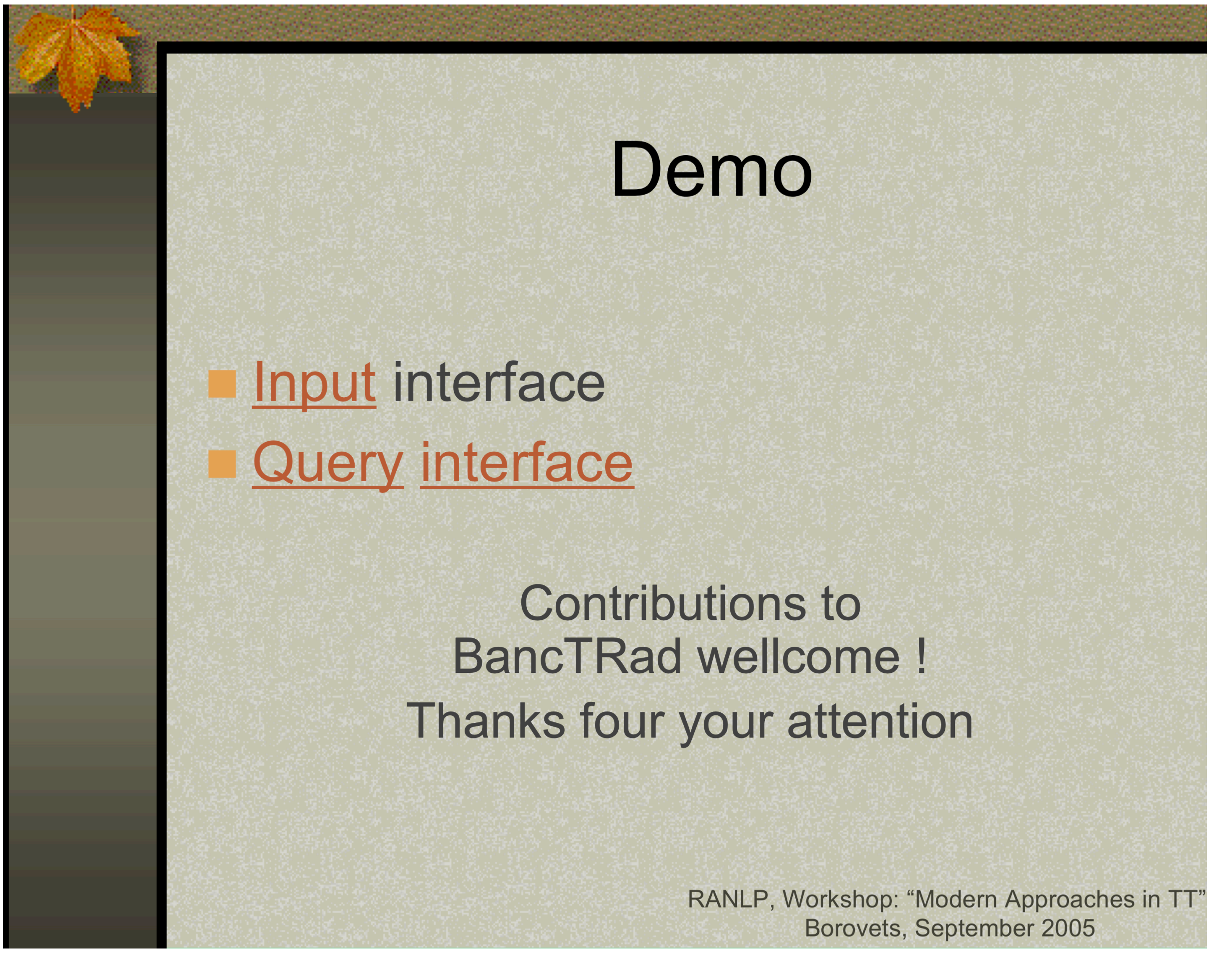
- makes the query processing
  - the CGI interfaces with information server
  - CQP allows
    - powerful query setting
    - access to linguistic and structural tags
    - aligned corpus querying

architecture (III): the search machine

- open access and platform independent →HTML-based interface
- user friendly/adaptable → two levels of expertise
  - basic mode: sequences of word forms
  - advanced mode: sequences of five quadruples (form, lemma, pos, and synt. function), negation
  - restrictions on extralinguistic features

Demo

- Input interface
- Query interface



Demo

- Input interface
- Query interface

Contributions to
BancTRad wellcome !

Thanks four your attention